

AUTOMATIC CONSTRUCTION OF DOMAIN-SPECIFIC CONCEPT STRUCTURES



Vom Fachbereich Informatik der Technischen Universität Darmstadt genehmigte

Dissertation

Zur Erlangung des akademischen Grades eines Doktor-Ingenieurs (Dr.-Ing.)

Von

Libo Chen

(Diplom Wirtschaftsinformatik)

geboren in Peking, China

Referent: Prof. Dr. Erich J. Neuhold

Koreferent: Prof. Dr. Thomas Hofmann

Tag der Einreichung: 09. 02. 2006

Tag der mündlichen Prüfung: 28. 03. 2006

Darmstadt 2006

D17

Darmstädter Dissertation

ABSTRACT

One of the greatest challenges for search engines and other search tools, which are developed to cope with the information overload, is the vocabulary mismatch problem, referring to the fact that different people usually use different vocabularies to describe the same concepts. This problem can first of all lead to unsatisfactory search results, because the keywords in search queries often do not match the indices of search engines – either the queries are too imprecise to describe users’ actual needs, or, although correctly formulated, the queries simply do not contain the keywords with which authors write their documents.

There is therefore a clear need to quickly build a concept structure for each possible topic or knowledge domain of user interest, which includes the most important concepts of a specific knowledge domain and the relationships between the concepts. Such concept structures can serve to standardize vocabularies in various knowledge domains, and help to bridge the vocabulary gap between information users, information creators, and search engines.

Since manual approaches often suffer from the problem of low coverage and high expense, this dissertation focuses on corpus based statistical approaches to automatically build domain-specific concept structures. These automatic approaches first select suitable text corpora to represent domains of interest, then find statistical evidence about terms in the text corpora, and finally perform statistical analysis upon the evidence to construct concept structures.

There exist two main challenges in the process of automatic construction of domain-specific concept structures: First, how the concepts in a domain can be found and extracted from text corpora (we refer to all important terms in a domain as concepts). Second, how the relationships between these concepts can be effectively determined.

For the task of concept extraction, we first introduce a notion of topicality to define the importance of a term, indicating how topical a term is to a specific domain. We further divide term topicality into two factors: term representativeness which indicates how well a term is capable of covering the topic area of a domain, and term specificity which indicates how specific a term is to a certain domain compared to other knowledge domains. We further present a novel approach for specificity calculation, where we not only collect information for the domain of interest, but also collect information for a set of reference domains. A statistical measure called the “Distribution Grade” is developed to compare the distribution of a term in different domains

to calculate its specificity more accurately. By combining representativeness and specificity, we are able to weight and sort terms in a text corpus according to their topicalities, and choose a limited number of top ranked terms as concepts in a domain of interest.

Relationship determination between concepts is usually based on a notion of common context of concepts, which is quantified by means of a similarity measure that compares the individual context of concepts with their common context. In this work, we first provide formal definitions and a detail analysis on two kinds of existing context – with one of them counting the frequency of co-occurrences of concepts in texts, and another considering the terms occurring in the neighbourhood of the concepts. We further introduce a new notion of context to overcome the limitations of previous approaches by combining evidence on both co-occurrences and neighbourhood terms. A mutual conditional probability model is presented as a general framework for formalizing the most successful similarity measures. Each type of context is then quantified by the probability model and combined to form a hybrid similarity measure to determine a “Generally Related” relationship. In addition, we also investigate the possibility of determining a “Broader/Narrower” relationship which plays an important role for building hierarchical concept structures. We show that considering the individual conditional probabilities in the mutual conditional probability model on the premise of a close “Generally Related” relationship helps to better find the “Broader/Narrower” relationship.

For an automatic evaluation of our approach, we employ widely accepted and manually built concept structures as “gold standards”, and automatically compare the extracted concepts and relationships with the entries in the gold standards. Experimental results show that our approaches achieve the best performance for a wide range of candidate terms and relationships, and for different types of text collections.

ZUSAMMENFASSUNG

Eine der größten Herausforderungen für Suchmaschinen und andere Suchwerkzeuge, die zur Bewältigung der hohen Informationsbelastung entwickelt werden, ist das Problem des Vokabularunterschieds (Vocabulary mismatch), was bedeutet, dass unterschiedliche Leute dazu tendieren, die gleichen Konzepte mit unterschiedlichen Termen zu beschreiben. Dieses Problem kann vor allem zu unbefriedigenden Suchergebnissen führen, da die Schlüsselwörter in Suchanfragen in vielen Fällen nicht mit den Einträgen von Suchmaschinenindizes übereinstimmen – entweder sind die Suchanfragen zu unpräzise, oder die Suchanfragen sind zwar inhaltlich richtig, aber sie enthalten nicht diejenigen Terme, mit denen Autoren ihre Texte formulieren.

Es ist deswegen notwendig, für jedes mögliche Themen- oder Wissensgebiet der Benutzerinteressen mit geringen Kosten eine Konzeptstruktur aufzubauen, die die wichtigsten Konzepte einer spezifischen Wissensdomäne und die Beziehungen zwischen den Konzepten umfasst. Solche Konzeptstrukturen können vor allem dazu dienen, das Vokabular in unterschiedlichen Wissensdomänen zu standardisieren und den Vokabularunterschied zwischen Informationsbenutzern, Informationserzeugern und Suchmaschinen zu überbrücken.

Da manuelle Methoden häufig unter den Problemen von niedrigem Abdeckungsgrad und hohen Kosten leiden, konzentriert sich diese Arbeit auf korpusbasierte statistische Verfahren für den automatischen Aufbau von domänenspezifischen Konzeptstrukturen. Diese automatischen Verfahren wählen zuerst passende Textkorpora zur Repräsentation der Zieldomänen, die für die Benutzer von Interesse sind. Aus diesen Textkorpora werden statistische Daten über das Auftreten der Terme ermittelt. Statistische Analysen werden schließlich auf der Basis dieser Daten durchgeführt, um Konzeptstrukturen zu konstruieren.

Es bestehen zwei Hauptherausforderungen bei dem automatischen Aufbau von domänenspezifischen Konzeptstrukturen: Erstens, wie die Konzepte einer Domäne aus den Textkorpora extrahiert werden können (wobei alle wichtigen Terme in einer Domäne als „Konzepte“ bezeichnet werden). Zweitens, wie die Beziehungen zwischen diesen Konzepten effektiv bestimmt werden können.

Für die Aufgabe der Konzeptextraktion führen wir zuerst einen Begriff von Topikalität ein, um die Wichtigkeit eines Terms zu definieren, die angibt, wie topikalisch ein Term zu einer

Domäne ist. Wir teilen weiterhin die Topikalität in zwei Faktoren ein: die Repräsentativität, die angibt, wie gut ein Term dazu fähig ist, den Themenbereich einer Domäne abzudecken; und die Spezifität, die angibt, wie spezifisch ein Term zu einer bestimmten Domäne ist im Vergleich zu anderen Wissensdomänen. Ein neues Verfahren zur Kalkulation von Spezifität wird entwickelt, wobei nicht nur die Information für die Zieldomäne, sondern auch die Information für eine Menge von Referenzdomänen berücksichtigt wird. Ein statistisches Maß – der "Verteilungsgrad" – wird entwickelt, um die Verteilungen eines Terms in den unterschiedlichen Domänen zu vergleichen, so dass die Spezifität des Terms genauer berechnet werden kann. Schließlich werden die Terme in einem Korpus nach ihren Topikalitäten gewichtet und sortiert. Eine begrenzte Anzahl von hochrangigen Termen wird als Konzepte ausgewählt.

Die Beziehung zwischen zwei Konzepten wird normalerweise durch ein Ähnlichkeitsmaß berechnet, das die Kontexte der einzelnen Konzepte mit ihrem gemeinsamen Kontext vergleicht. Wir geben in dieser Arbeit zuerst formale Definitionen und eine Detailanalyse von zwei existierenden Kontexttypen. Bei dem einen Kontexttyp wird die Häufigkeit des gemeinsamen Auftretens von Konzepten in den Texten ermittelt, während bei dem anderen die Nachbarterme in der Umgebung von Konzepten berücksichtigt werden. Wir führen weiterhin eine neue Art von Kontext ein, um die Beschränkungen der existierenden Kontexttypen zu überwinden, wobei sowohl gemeinsames Auftreten als auch Nachbarterme berücksichtigt werden. Ein Modell der gegenseitig bedingten Wahrscheinlichkeit wird als ein allgemeiner Rahmen für die Formalisierung der erfolgreichsten Ähnlichkeitsmaße vorgestellt. Jeder Kontexttyp wird dann durch das Wahrscheinlichkeitsmodell quantitativ bestimmt und kombiniert, um ein hybrides Ähnlichkeitsmaß für die Bestimmung einer "Allgemein verwandt" -Beziehung zu bilden. Zusätzlich suchen wir nach einer Möglichkeit, die "Ober-Unterkonzept" -Beziehung zu bestimmen, die eine wichtige Rolle für die Konstruktion hierarchischer Konzeptstrukturen spielt. Wie wir feststellen können, lässt sich die "Ober-Unterkonzept" -Beziehung besser berechnen, wenn die einzelnen bedingten Wahrscheinlichkeiten in dem Modell der gegenseitig bedingten Wahrscheinlichkeit auf der Basis einer engen "Allgemein verwandt" -Beziehung berücksichtigt werden.

Für eine automatische Evaluation unserer Verfahren setzen wir bekannte und manuell aufgebaute Konzeptstrukturen ein, die als „Goldstandards“ bezeichnet werden. Wir vergleichen automatisch die extrahierten Konzepte und Beziehungen mit den Einträgen in den Goldstandards.

Experimentelle Ergebnisse zeigen, dass unsere Verfahren die beste Performanz für eine große Breite von Kandidatermen/Kandidatenbeziehungen und für unterschiedliche Datenkollektionen liefern.

ACKNOWLEDGEMENTS

First of all I would like to express my deepest gratitude to my supervisor, Prof. Dr. Erich J. Neuhold, for his guidance during the development of this work, and without whose help this thesis would not have been possible. I would like to thank my second supervisor Prof. Dr. Thomas Hofmann for his valuable advice and suggestions on this thesis. My thanks also go to other members of my doctoral examination committee: Prof. Dr. Karsten Weihe (the chair), Prof. Dr. José Luis Encarnação, and Prof. Dr. Thomas Kühne.

I would also like to take this opportunity to thank Dr. Thomas Kamps for his great support and mentoring during my Ph.D. study. Special acknowledgement is given to Dr. Ulrich Thiel for his patience, encouragement, and those innumerable lessons I learned from him throughout my study. I will also give a special thank to Dr. Peter Fankhauser, who provided valuable suggestions and comments that improved the contents and presentation of this dissertation to a great extent. This work was accomplished with the financial and technical support of Fraunhofer IPSI. I sincerely appreciate all the kind help of the colleagues in the institute. Thanks are also due to the support of the Computer Science Department, Darmstadt University of Technology.

Finally, I would like to thank my wife Xiujuan and my parents for their love and dedication which was in the end what made my research possible.

LIST OF CONTENT

ABSTRACT.....	I
ZUSAMMENFASSUNG.....	III
ACKNOWLEDGEMENTS	VI
CHAPTER 1 INTRODUCTION.....	1
1.1 THE EXPLOSIVE GROWTH OF INFORMATION	1
1.2 THE PROBLEM OF VOCABULARY MISMATCH	2
1.3 DOMAIN-SPECIFIC CONCEPT STRUCTURES.....	4
1.4 AUTOMATIC CONSTRUCTION OF DOMAIN-SPECIFIC CONCEPT STRUCTURES	8
1.5 CORPUS BASED STATISTICAL APPROACHES.....	9
1.6 RESEARCH CONTRIBUTIONS	13
1.7 ORGANIZATION OF THE DISSERTATION.....	15
CHAPTER 2 RELATED WORK.....	16
2.1 SOURCE SELECTION.....	16
2.1.1 Conventional Text Corpora.....	16
2.1.2 Search Results of Web Search Engines	17
2.1.3 Web Directories	18
2.2 TERM WEIGHTING.....	20
2.2.1 Statistical Approaches.....	20
2.2.2 Linguistic Approaches	24
2.3 RELATIONSHIP DETERMINATION.....	26
2.3.1 Statistical Approaches.....	26
2.3.2 Linguistic Approaches	27
2.3.3 Non-corpus-based Approaches	28
2.4 EVALUATION METHODS	29
2.4.1 Manual Evaluation	29

2.4.2 Automatic Evaluation	30
CHAPTER 3 AUTOMATIC EXTRACTION OF DOMAIN-TOPICAL	
CONCEPTS.....	33
3.1 WEB DIRECTORIES AS A SOURCE FOR COLLECTING DOMAIN-	
TOPICAL CONCEPTS	33
3.2 TOPICALITY BASED TERM WEIGHTING.....	35
3.2.1 Computing Term Topicality	35
3.2.2 F_d/F_c , Odds-Ratio and Kullback-Leibler Divergence	36
3.2.3 Distributional Topicality Weighting	39
3.3 AUTOMATIC EVALUATION.....	44
3.3.1 Principle of Automatic Evaluation.....	45
3.3.2 Experimental Setup.....	49
3.3.3 Evaluating the Quality of Web Directories as a Source	51
3.3.4 Evaluating Weighting Approaches for Concept Extraction.....	54
3.3.5 Automatic Evaluation on a Second Domain	59
3.4 MANUAL EVALUATION AND EXAMPLES	61
3.5 CONCLUSION	63
CHAPTER 4 STATISTICAL RELATIONSHIP DETERMINATION	65
4.1 CONDITIONAL PROBABILITY MODEL	66
4.2 OCCURRENCE CONTEXT	67
4.2.1 Formal Definitions	67
4.2.2 Similarity Measures Based on Occurrence Context	69
4.2.3 Problems of Occurrence Based Approaches.....	74
4.3 CONTENT CONTEXT	76
4.3.1 Formal Definitions	76
4.3.2 Similarity Measures Based on Content Context	77
4.3.3 Problems of Content Based Approaches.....	79
4.4 COMBINING OCCURRENCE CONTEXT WITH CONTENT CONTEXT ...	83
4.4.1 Comparison with Purely Content Based Common Context	83

4.4.2	Comparison with Purely Occurrence Based Common Context.....	86
4.4.3	A Combined Similarity Measure	88
4.5	DETERMINATION OF THE “BROADER/NARROWER” RELATIONSHIP	89
4.6	AUTOMATIC EVALUATION.....	93
4.6.1	The “Generally Related” Relationship.....	95
4.6.2	The “Broader/Narrower” Relationship	99
4.6.3	Automatic Evaluation in a Second Domain.....	108
4.7	MANUAL EVALUATION AND EXAMPLES	111
4.7.1	rel_sqrt_occ_con.....	111
4.7.2	Aspect Ratio.....	116
4.7.3	The “Broader/Narrower” Relationship	119
4.8	CONCLUSION.....	121
CHAPTER 5	CONCLUSTION	122
5.1	SUMMARY OF THE THESIS	122
5.2	FUTURE WORK.....	123
REFERENCES.....		125
CURRICULUM VITAE.....		133

LIST OF FIGURES

Figure 1.1. The problem of vocabulary mismatch	3
Figure 1.2. A small part of the astronomy thesaurus	7
Figure 3.1. Domain distribution of Term 1 and Term 2	42
Figure 3.2. Theoretical F-measure curves of different weighting methods	48
Figure 3.3. F-measure curves for different weighting methods – Astronomy / Data basis 1	55
Figure 3.4. F-measure curves in a smaller rang – Astronomy / Data basis 1	56
Figure 3.5. F-measure curves – Astronomy / Data basis 2	57
Figure 3.6. F-measure curves in a smaller range – Astronomy / Data basis 2	58
Figure 3.7. F-measure curves for different weighting methods – Construction	60
Figure 4.1. Relationship determination through the common context terms	76
Figure 4.2. Topics involving “Sun” and “Photograph”	82
Figure 4.3. Different types of context of common context	84
Figure 4.4. The relationship between a general term and a relatively specific term	86
Figure 4.5. The distributions of “Linux” and “SUSE” in different text segments	91
Figure 4.6. Comparison of different variations of PMI	96
Figure 4.7. Comparison of other similarity measures	97
Figure 4.8. Comparison of different similarity measures developed in this work	98
Figure 4.9. Comparing the relative ability of different similarity approaches in finding bn relationships	101
Figure 4.10. Comparing the absolute ability of different similarity measures to find bn relationships	103
Figure 4.11. Combining rel_sqrt_occ and P_occ_doc	105
Figure 4.12. Combinations of $P(t_2 t_1)$ and $P(t_1 t_2)$	107
Figure 4.13. Comparison of different similarity measures in a second domain	109
Figure 4.14. Relative abilities of similarity measures in finding bn relationships	110
Figure 4.15. Absolute abilities of similarity measures in finding bn relationships	111

LIST OF TABLES

Table 3.1. Only using “keywords” and “description” – Web directories	52
Table 3.2. Using the whole page – Web directories	52
Table 4.1. Contingency table for combinations of term pairs t_1 and t_2	70
Table 4.2. Similarity measures.....	71
Table 4.3. Distributional similarity measures	78
Table 4.4. Different notions of context and common context for calculating PMI	95
Table 4.5. Statistics of S_1 and S_2	112
Table 4.6. Top 30 Relationships in S_1 - S_2 and S_2 - S_1	113
Table 4.7. Statistics of S_3 compared with S_1 and S_2	116
Table 4.8. Top 30 relationships in S_3 - S_2 - S_1	117
Table 4.9. Top 60 Relationships weighted by combination of conditional probabilities.....	120

CHAPTER 1 INTRODUCTION

In this chapter, we first describe two major problems in information processing – information overload and vocabulary mismatch, in Sections 1.1 and 1.2 respectively. In Section 1.3, we introduce domain-specific concept structures as a solution to the problems. In Section 1.4, we discuss the necessity and possibility of automatically building domain-specific concept structures. We then introduce corpus-based statistical approaches as one of the most important automatic approaches in Section 1.5. Section 1.6 summarizes the research contribution of this work, and Section 1.7 describes the organization of the remainder of the thesis.

1.1 THE EXPLOSIVE GROWTH OF INFORMATION

With the tremendous development of computer technology and the emergence of the Internet, the amount of digital information has been growing explosively in the last years. The drawback of this high information availability is that people are easily overwhelmed by the huge amount of information, and are not able to find the useful information they actually need.

Search engines prevailing on the World Wide Web (WWW) and many local networks are one of the means to help users out of the information jungle. A search engine usually indexes documents, compares the terms in documents with user queries, and provides users with a set of documents as search result ranked by some algorithms.

Another way to solve the problem of information overload is to classify information into various knowledge domains, i.e. knowledge areas concerning different specific topics, such as science, politics, sports, arts etc.. Larger domains may further contain smaller domains, e.g. the domain of science can be further divided into physics, chemistry, mathematics, biology etc.. Classifying information in domains has been used

as a means of organising information for a long time, like the Decimal Classification System developed by Melvill Dewey in 1873 [63], which has been widely applied in libraries to classify books. More recent attempts of using domains for classifying information are web directory systems on the WWW, such as Yahoo!, Dmoz and Google Directory (a version based on Dmoz), where human experts manually classify web documents into different pre-defined domains and sub-domains. The information contained in a specific domain is expected to be more precise and of higher quality compared to the information on an unorganised basis, because by classification, relevant information with respect to the topic of a domain is included and irrelevant information is excluded. Users will have a better chance of finding the information they are interested in by restricting their search to a specific topic domain.

1.2 THE PROBLEM OF VOCABULARY MISMATCH

Besides the information overload, another challenge in information processing is the vocabulary mismatch problem, referring to the fact that people tend to use different terms to describe a concept. Due to their different backgrounds and expertise, the chance that two people use the same term to describe a concept is quite low, and because of the learning process and the evolution of concepts, even the same person may use different terms to describe the same concept at different times [35][7].

Figure 1.1 depicts the complexity of the problem of vocabulary mismatch. In this Figure, we classify people engaged in information processing tasks into two groups: information creators, who produce information, and information users, who search for information. The vocabulary mismatch problem can be identified both within and between the individual groups.

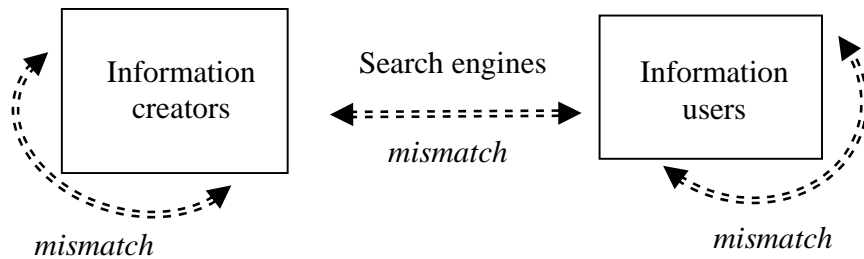


Figure 1.1. The problem of vocabulary mismatch

For example, an information creator, who is an expert in Information Retrieval (IR), may use the term “Boolean information retrieval system” to describe a special kind of search engines that connect terms in a query with Boolean operators AND, OR, or NOT to refine the search scope. A novice in IR, however, who is not aware of the existence of other retrieval forms, may simply refer to them as “search engine”. An outsider to IR may even not know the concept “search engine” and use “Google” – the name of one of the most popular search engines to refer to all kinds of search engines. Additionally, people at the same level of expertise are also likely to describe similar objects differently. For example, some people tend to use the term “name” to tag the name of a book author on a web page in XML, while other prefer the term “author” for the annotation.

Similar phenomena can be observed with information users in searching tasks, where the search queries formulated by different users often vary to a great extent, although they aim to find a same topic.

The vocabulary mismatch problem between information creators and information users is more severe. On one hand, compared to information creators, information users tend to have less expertise and patience to formulate precise queries. Previous research [84] has shown that the average length of web queries is less than two words, which usually do not contain sufficient information to cover necessary search terms. On the other hand, even a precisely formulated user query may not guarantee a successful search, because it may simply not contain the keywords with which information creators write

their documents. In both cases, the queries of users will not match the indices of search engines, which are actually indexed vocabularies of information creators. This means that, if a user searching for a topic does not exactly know how information creators describe the topic in their documents, he will not be able to find information about this topic using search engines. How to bridge the vocabulary gap between information searchers and information creators is one of the greatest technical challenges that the modern search engines such as Google and Altavista have to be confronted with.

1.3 DOMAIN-SPECIFIC CONCEPT STRUCTURES

Designed for vocabulary control and knowledge representation, a concept structure is a systematic organization of important vocabularies, which usually contains a finite set of carefully chosen concepts and terminologies, and some kinds of relationships between these concepts and terminologies. According to the relationships, the vocabularies may be further organized in structures of different forms for easier access.

The most common concept structures include taxonomies, thesauri and ontologies. Since a precise definition for each type of concept structure is not the focus of this work – with the exact definition for ontology in the scope of computer science still being a debated issue, we only give informal descriptions below to provide an intuitive overview. A more detailed discussion about the definition and evolution of different types of concept structures can be found in [37].

While taxonomies are usually hierarchical, and only contain the broader/narrower relationship, thesauri extend this with the related and synonym relationships and the Scope Note, which are used to clarify the exact meaning of a term [45]. An ontology applied in the scope of computer science can also integrate inference rules, and model unlimited kinds of relationships between concepts [6][42]. In our work, we will focus on the abstract view of a concept structure in containing important concepts and

relationships in a domain, without considering the detailed difference between various types of concept structures existing in actuality.

According to different scopes of coverage, concept structures can be roughly classified into two groups: general-purpose and domain-specific. General-purpose concept structures like Roget's thesaurus¹ and Wordnet² usually model general linguistic usage of words in a language, and may contain lexical/semantic relationships between words such as synonym, antonym, hypernym/hyponym, holonym/meronym etc. They are valuable tools in computational linguistics such as statistical model smoothing and word sense disambiguation.

In many other tasks, however, people are more interested in specific information from a certain domain. As mentioned in Section 1.1.1, people tend to classify information into domains to avoid the heavy mental burden of processing all information at the same time. So the information need of a user when performing a searching task is normally explicitly or implicitly restricted to a specific domain. The same applies to an author (i.e. an information creator) when creating a document. A general-purpose concept structure is usually not capable of providing satisfactory coverage and depth in a domain with respect to both vocabularies and relationships. In contrast, domain-specific concept structures are constructed separately in different individual domains, aiming to cover the knowledge of the individual domain as comprehensively and as precisely as possible. They are therefore capable of including many more high quality concepts and relationships in a domain.

Domain-specific concept structures can be used as a powerful tool to conquer the vocabulary mismatch problem mentioned in the last section.

¹ <http://thesaurus.reference.com/>

² <http://wordnet.princeton.edu/obtain>

Information creators can for example reach an agreement to use a well constructed and widely accepted domain-specific concept structure as a standard of vocabulary when creating documents in a domain. This helps facilitate a more effective communication between different people and between different machine agents that automatically perform intelligent tasks for human beings.

For overcoming the vocabulary mismatch problem between information creators and information users, a search engine may suggest to users more appropriate keywords with the help of domain-specific concept structures that are consistent with the vocabularies of the creators, if the user queries are too vague or inaccurate. User queries can be either automatically modified by the suggested terms, which is usually referred to as automatic query expansion; or users are required to manually select the desirable keywords, referred to as interactive query expansion.

In an example of interactive query expansion, we suppose a user is fascinated with the beautiful light “pearl” at the edge of the sun during a solar eclipse and wants to know more about this amazing astronomical phenomenon. Since the user does not know the exact keywords for it, he simply inputs the term “eclipse” into a search engine and is likely to be disappointed by thousands of documents in the search results that do not seem to have anything to do with what he is looking for, even the search is already restricted to the domain of astronomy. To solve this problem, a search engine with an integrated astronomy concept structure may provide the user a small part of the concept structure related to the topic eclipse in a certain form as shown in Figure 1.2.

Depending on their domain expertise, different users may use this concept structure in different ways. Those who have sufficient domain knowledge tend to know the meaning of the concepts in the structure. As they cannot come up with the right keywords for searching, they may simply browse through this concept structure, following the pre-defined relationships from one important terminology to another until the correct keywords “Bailys bead” and “Diamond ring effect” are found. Other users who do not

know the domain very well will have to either consult the information in the Scope Notes or any documents and pictures attached to the individual concepts for their interpretations, or input the corresponding concepts into search engines to find out the meanings of these concepts. In this way, the users get gradually educated with the relevant domain knowledge, until they find the right keywords.

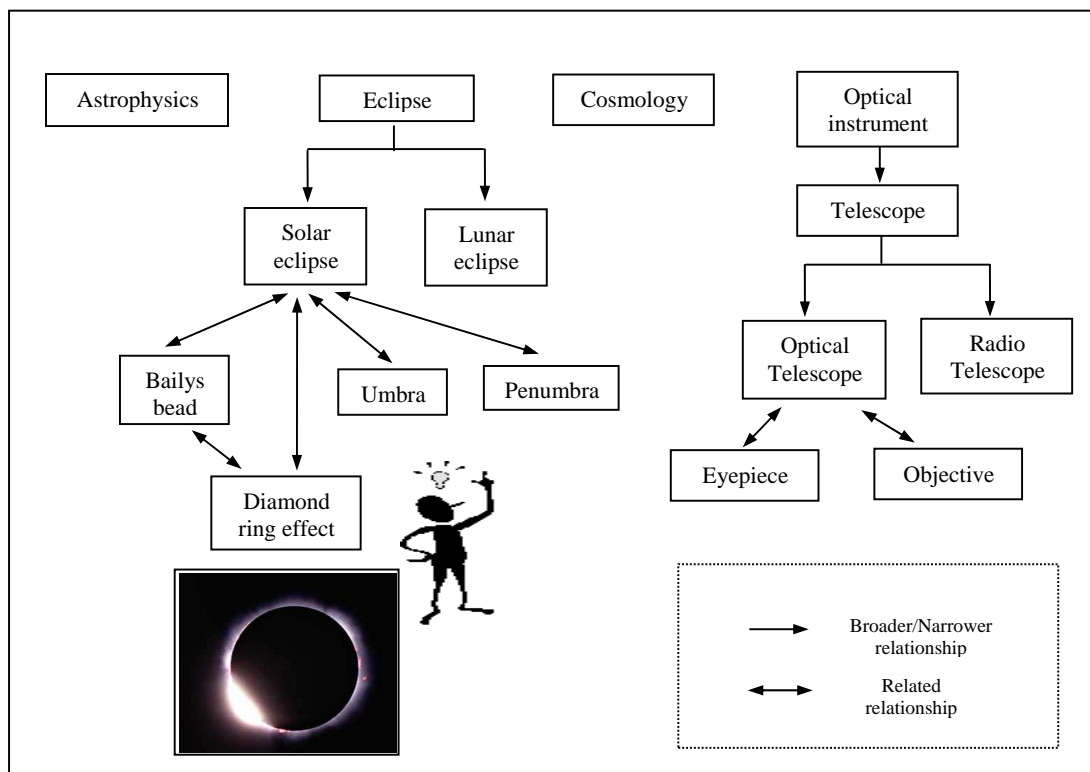


Figure 1.2. A small part of the astronomy thesaurus

It is worth noting that for the purpose of educating novice users with domain knowledge, domain-specific concept structures are also often used independently from a query expansion task. That is, a user jumps into a concept structure and makes a “memory jogging” across it without aiming to find a certain keyword. The concept structures serve here as a skeleton of knowledge, which help users to learn domain knowledge more systematically and more quickly.

Another possible application of domain-specific concept structures is document classification, where different documents are assigned to the appropriate concepts in a concept structure by considering the conformance between the documents and the concepts. When assigned appropriately, the documents attached to a concept can serve as an interpretation of this concept, which, as mentioned above, can be very useful in helping novice users to learn domain knowledge when they browse a concept structure.

1.4 AUTOMATIC CONSTRUCTION OF DOMAIN-SPECIFIC CONCEPT STRUCTURES

A domain-specific concept structure can be built either manually or automatically. In a typical process of manual construction, for example, subject experts are involved in defining the boundaries of the subject area. They then have to determine terms for the defined area, based on their own expertise and a variety of relevant sources that could be found, such as indexes, encyclopedias, handbooks, textbooks, journal titles and abstracts, as well as any existing and relevant thesauri or vocabulary systems. After the terms are identified, each term is analysed for its related vocabulary, including synonyms, broader/narrower terms and related terms, and sometimes also definitions and scope notes of terms.

Although the manual process guarantees the quality of the concept structures to some extent, it is usually rather expensive and time consuming, and suffers from the problem of low coverage: on one hand, only a very small part of a domain – usually the most important top level concepts – can be covered due to the resource limitation, on the other hand, it would be impossible to manually construct concept structures for all potentially interesting topics in a language. In addition, after a concept structure is built, it must be constantly updated to reflect the change of information within the area, which often cannot be afforded by the slow response times of manual process. There is a clear

need for fully automatic methods to replace or at least to support manual creation and updating of domain specific concept structures.

The procedures deployed in automatic construction roughly resemble those in manual construction. First, suitable domain specific text corpora or existing concept structures are chosen as a basis for processing. In the second step, important concepts and relationships are automatically determined by computer algorithms that take various types of statistical or linguistic evidence into consideration. Since no human intervention is required, the whole process usually takes only some hours or several days instead of years of work in manual construction.

Due to the low cost and high flexibility, automatic approaches are capable of processing extensive amount of domain specific data and of quickly reacting to information change. This enables an easy construction and maintenance of concept structures for each (potential) topic of user interest.

The problem of automatic approaches is its relatively low quality. Compared to manually constructed concept structures, automatically built ones often contain “noise”, i.e. irrelevant concepts and relationships. In addition, it is generally difficult to distinguish different types of relationships, e.g. “Related” and “Broader/Narrower” by using automatic methods, which are typically contained in a manually built thesaurus. It is therefore crucial to develop automatic algorithms that optimally build concept structures with minimal noise.

1.5 CORPUS BASED STATISTICAL APPROACHES

As one of the most important automatic approaches for constructing concept structures, corpus based statistical approaches do not assume the existence of any pre-defined concept structures and use solely text corpora as sources. Based on carefully chosen text corpora with good domain representation, these approaches can get access to

a wide range of domain-specific information, and collect sufficient statistical evidence, such as term distributions among documents/domains or co-occurrence information of two terms in different text segments, for the automatic construction.

In contrast to purely statistical approaches, other approaches also take linguistic evidence into consideration, which, however, requires profound linguistic knowledge and well performing natural language processing (NLP) tools to determine syntactic roles of terms or discover linguistic patterns in texts, as shown in the work of [42][53]. As both fields of statistical and linguistic analysis are considerably comprehensive, we mainly deal with purely statistical methods in this thesis for a better focus.

Since it is generally difficult to distinguish between abstract concepts and instances in a text corpus by applying purely statistical analysis, statistical approaches usually regard all important terms in a domain as potential concepts [46][67][13][56]. Consequently, proper nouns like the name of a person, an organisation, or a special device can also be taken as concepts, which are especially useful in query modification tasks in IR, because users with less domain knowledge sometimes begin their search with quite concrete terms like “Google”, rather than using abstract concepts like “information retrieval system”.

There are two main challenges for corpus based statistical approaches:

1. Concept extraction.

A domain-specific text corpus with sufficient coverage usually contains millions or tens of millions of terms. With noise and less important terms being filtered out, only a restricted number of the most important terms should be extracted and included in concept structures. In practice, this number may vary to a great extent depending on the particular demands of different users and applications, usually ranging between thousands and tens of thousands.

The key questions here are: how is the importance of a term in a domain defined? Which statistical evidence in text corpora can be used to reflect term importance,

and how can automatic mechanisms be developed to quantify the statistical evidence for an effective calculation of term importance?

Traditional term weighting approaches such as TF-IDF [73], which has been successfully used in Information Retrieval tasks, and feature selection [88][62] methods, usually applied for text classification, are not suitable for the task at hand, because they define the importance of a term according to its ability to distinguish documents or categories, not its ability to indicate the topic of a domain. Several term extraction approaches for building concept structures [76][38][50] only consider incomplete evidence to calculate term importance. More advanced techniques should be developed to achieve better performance for concept extraction.

2. Relationship determination

After salient concepts in a domain are extracted, the relationships between the concepts should be determined by considering statistical evidence in text corpora. As in concept extraction, only a limited number of the most important relationships can be included in final concept structures.

The key issues are to determine and qualify different kinds of statistical evidence for an optimal relationship determination. It is also desirable to distinguish different relationship types as in a manually built concept structure.

To this end, there exist numerous corpus-based approaches that use occurrence frequency of terms in one or the other form. Traditional approaches based on co-occurrence of terms [73][40] ignore the actual content in the context of a co-occurrence. More recent approaches [28][36][80] that take into account the actual content in context do not distinguish between significant content, which contains important information for relationship determination, and spurious content, which contains only noise and should actually be ignored. Researches as in [76] which attempt to build hierarchical concept structures, do not show

promising results in finding the “Broader/Narrower” relationship either. There is a clear need for a systematic analysis of the numerous previous approaches and new approaches should be developed for a more effective relationship determination.

In addition, there exist two secondary problems that should also be addressed.

3. Source selection

Choosing an appropriate source as the data basis is the first step in corpus based approaches. A good domain-specific source should have sufficient domain coverage, contain as little noise as possible, be flexible to changes and easily accessible. Traditional text corpora that have been widely applied in Information Retrieval and NLP tasks usually suffer from the problem of data sparseness, meaning that they do not contain enough important terms and co-occurrence information between terms for a domain [47] [41]. In contrast, WWW resources, especially the web directories, are capable of providing considerably more domain information than traditional text corpora, and can therefore also be considered as a valuable source for text processing tasks.

4. Evaluation

Evaluating the quality of a concept structure is usually difficult. While manual evaluation suffers from the problems of high cost and assessment errors, automatic evaluation methods should be developed to effectively deal with the numerous terms and relationships that are automatically extracted from text corpora.

1.6 RESEARCH CONTRIBUTIONS

This dissertation focuses on automatic construction of domain-specific concept structures using corpus-based statistical approaches. Two main contributions are made in this research:

1. For the task of concept extraction, we introduce a notion of *topicality* to indicate the importance of a term in a target domain. We find that the topicality of a term is actually a combination of two factors: *representativeness* and *specificity*. While representativeness measures how well a term is capable of representing a domain, specificity indicates how specific a term is in the target domain compared to other domains. Most of the previous research such as [76][38] implicitly uses either representativeness or specificity for calculating the importance of a term in the target domain. A few topicality based weighting approaches like [50] do not explicitly distinguish between representativeness and specificity. Further, we present a novel approach for calculating term specificity. In contrast to previous methods, which compare the distribution of a term in the target domain with its distribution in a much larger text collection representing the whole language, we further divide the larger text collection into different reference domains, and introduce a measure called the “*Distribution Grade*” to compare the distributions of target terms in different domains, so that a more accurate specificity calculation can be achieved. By combining representativeness and specificity, we are able to weight and sort terms in a text corpus according to their topicalities, and choose a limited number of top ranked terms as concepts in a domain of interest. Experimental results in Chapter 3 show that our methods achieve the best performance for concept extraction for a wide range of candidate terms on different types of data basis.

2. For relationship determination, we present a *mutual conditional probability model* as a general framework for formalizing the most successful similarity measures to determine a “Generally Related” relationship. Independent from similarity measures, we provide formal definitions and a detailed analysis on two kinds of frequently applied context type – the *occurrence context* and the *content context*. While the occurrence context of terms is defined as the text segments within which they occur individually and common content context as the text segments within which they co-occur, the content context is defined as the content of text segments around individual terms and common context as the intersection of the content of these contexts. Because both context types have individual strengths and weaknesses, we suggest a *new notion of common context* to overcome their limitations, which requires the terms co-occurring in the same text segments, and at the same time considers the content in the common text segments. Each type of context and common context is quantified by means of conditional probabilities and a *hybrid similarity measure* is formed, which conjunctively combines the evidence provided by each notion of context. In addition, we also show that considering the individual conditional probabilities in the mutual conditional probability model on the premise of a close “Generally Related” relationship helps to better find a “*Broader/Narrower*” relationship. Experimental results confirm our analysis and show that the hybrid similarity measure provides the best performance in relationship determination by achieving an improvement of nearly 70% at the best F-measure value compared with a traditional document based co-occurrence analysis.

In addition to the two main contributions described above, we also demonstrate that web category systems like the Yahoo! and Google Directories are good sources for our task. A web category system usually has different directories and sub-directories, with

each of them containing numerous web sites that are relevant to a specific domain. For the task of automatically constructing concept structure in a domain, we first find the appropriate category in the category system; then use the whole category as a “bag” of web sites without considering its structure; finally, we crawl all of these web sites and use them as the base text corpus. In Chapter 3 we will show that such text corpora usually represent target domains well, and have sufficient coverage of domain-specific concepts and relationships.

In addition, we develop an automatic evaluation method, which employs widely accepted and manually-built concept structures as gold standards to automatically evaluate the quality of a text corpus and the quality of different approaches for concept extraction and relationship determination. For a more intuitive comparison, the F-measure [70] is adopted to combine precision and recall into one measure.

1.7 ORGANIZATION OF THE DISSERTATION

The remainder of this dissertation is organized in the following manner. In Chapter 2, we discuss related work with respect to source selection, term weighting, relationship determination and evaluation. Chapter 3 deals with concept extraction, where representativeness and specificity of a term are distinguished. A novel approach is developed to compute specificity more accurately. The representativeness and specificity of a term are finally combined to calculate topicality. In Chapter 4, we provide novel approaches for relationship determination by introducing and analysing different notions of context and common context. They are then quantified by a model of conditional probabilities and are finally combined to form a hybrid similarity measure. In this chapter, we also discuss the possibility of better finding the “Broader/Narrower” relationship. Chapter 5 draws conclusions and points to future work.

CHAPTER 2 RELATED WORK

In this chapter we review different methods for constructing domain-specific concept structures, where the emphasis is put on corpus based statistical approaches. In Section 2.1 we discuss relevant work for selecting suitable sources as the basis text corpora. Different approaches for concept extraction and relationship determination are reviewed in Section 2.2 and Section 2.3, respectively. Section 2.4 describes the most relevant evaluation methods.

2.1 SOURCE SELECTION

Selecting an appropriate source as the basis for further processing is a crucial task for all corpus-based approaches. Different kinds of sources have been applied in previous relevant research, including conventional text corpora, search results of web search engines and web directories.

2.1.1 Conventional Text Corpora

Before the emergence of the World Wide Web (WWW), text corpora were widely used as sources for various research fields, such as Information Retrieval (IR), natural language processing and automatic thesaurus construction. The most popular English text corpora are, for example, the Brown Corpus [9], the Reuters Corpus of news stories [69], TREC corpora [81], Penn Treebank [66], LDC corpora [53] and TDT corpora [79] etc.. For the purpose of automatic construction of domain-specific concept structures, domain specific corpora are usually applied, as in the work of [13], where four main sources of textual documents in an electronic community system were used for automatic thesaurus construction in the domain of molecular biology. These sources provided different

forums for community members to discuss topics ranging from finished project and ongoing research, to laboratory observations and personal communications, including a book, journal abstracts, a news letter and conference proceedings.

The main problem of using conventional text corpora as sources is the so-called “data sparseness” problem, which means that conventional text corpora usually do not contain enough interesting domain concepts and co-occurrence information between concepts, as shown in the work of Keller et. al. [47] and Grefenstette [41].

2.1.2 Search Results of Web Search Engines

The WWW is another possible resource for text processing, which has been demonstrated to contain much more interesting domain information than traditional text corpora. Keller et. al. [47] showed in their paper that the WWW is capable of providing many bi-grams that are usually unseen in a conventional text corpus.

Chen et. al. [16] built web thesauri from web link structure. The first step of their approach is to find high quality and representative web sites for a domain. To do this, they submitted the domain name as a query to the Google directory to obtain a list of authority websites.

Liu et. al. [56] also submitted the name of a topic as a query to a web search engine and mined concepts from the top ranked search results provided by the search engine.

Turney [82] and Baroni et. al. [4] used the search results of web search engines for synonym detection, in the hope that the numerous web documents indexed by search engines could build a text collection that would have a less severe “Data sparseness” problem than a normal text corpus.

Grefenstette [41] explored the possibility of using the WWW as a source for machine translation tasks. He showed that using a text corpus (e.g. BNC for British National Corpus) for counting term frequency had caused an acute data sparseness problem, whereas using the WWW as a source led to a much better result.

However, Grefenstette also pointed out in the same paper that the drawback of using the WWW is that the information on the WWW is usually far more noisy than in a well-edited text corpus. In our work this problem is addressed in two ways: 1. Using a more reliable web source such as web directories. 2. Applying more effective weighting approaches for better filtering web noise. We will present these solutions in more detail in the next chapter.

2.1.3 Web Directories

Due to their relatively high accuracy and accessibility, web directories like Yahoo! directory³, Dmoz⁴ and Google directory⁵ (based on Dmoz) are valuable sources for text processing tasks.

Chen et. al. [12] used an early version of a web directory system LookSmart⁶ for document categorization, which contained 13 top-level categories. The documents originally contained in the web directories were used for training. The documents contained in search results were then automatically categorized to the LookSmart categories by using the Support Vector Machine technique.

Beitzel et. al. [5] used web directories to automatically evaluate the performance of an information retrieval system. The basic assumption was that documents contained in different human-edited web categories are highly relevant to the respective categories. Search results in responding to a search query were compared with the documents in the relevant web category, whose label is similar to the terms in the query. The larger the two sets of documents overlap, the better the search results will be regarded.

³ <http://dir.yahoo.com/>

⁴ <http://dmoz.org/>, Open Directory Project

⁵ <http://directory.google.com/>

⁶ <http://www.findarticles.com/>

An early work of Lewis and Croft [52] in 1990 took advantage of manually built categories in the Journal “Computing Reviews” for relationship determination between terms, where two terms were regarded to be related if they co-occur in same categories.

Glover et. al. [38] collected terms from 41 categories in Dmoz with each category representing a specific domain. All categories together formed a general document collection. By comparing the term frequency in a certain domain with that in the general document collection, Glover was able to distinguish terms such as “parent”, “self” and “child” – a term is a “self” if it appears commonly in the domain, but less commonly in the whole collection; a term is a “parent” if it appears both commonly in the domain and in the whole collection; a term is a “child” if it appears commonly in the domain and very rarely in the collection. In this way, Glover hoped to find the most important terms in a domain and build a term hierarchy with three layers, namely “parent”, “self” and “child”. A large overlap between the term hierarchy and the labels of subcategories in the corresponding category was shown in an evaluation.

It is worth noting that, in a web directory system, if a category is used to represent a domain, the labels of its sub-categories usually form a simple hierarchical term structure. It seems that this term structure can be directly used to construct a domain-specific concept structure. One of the problems of this idea is that the hierarchy of the subcategories is usually quite shallow and the labels of the subcategories often too general. In the Yahoo! directory for example, there are only a few (sub)categories corresponding to the domain of astronomy/telescope, such as “telescopes”, “Amateur”, “Telescope Making”, although it is a fairly important sub-domain of astronomy. A satisfying domain-specific concept structure is expected to include much more concepts and relationships, which can hardly be effectively captured by the label structure of a category system that is maintained by human experts. Another problem is that sub-categories are usually built for an effective organization of documents, not for representing the knowledge in a domain. The astronomy domain in Yahoo!, for example,

contains sub-categories such as “Companies”, “News and Media” and “ask an Expert”, which are not specific to astronomy, and thus should not be included in the concept structure. Finally, as indicated by Lawrie [50], the structure in a category system tends to be fairly static and not particularly adaptable to the rapid change of the web documents. We therefore do not consider the sub-category structure of a web category when constructing domain-specific concept structures.

2.2 TERM WEIGHTING

The goal of term weighting is to give a quantitative measure for computing the importance of a term. The definition of term importance may vary in different applications, so that a term which is important for one application (such as Information Retrieval) may be regarded as unimportant in another application (such as concept extraction). The numerous existing term weighting approaches can be roughly divided into two groups, statistical approaches and linguistic approaches, depending on whether statistical or linguistic evidence is taken into consideration. In the following we focus on statistical approaches which are more closely related to our work. A brief introduction of several relevant linguistic approaches is provided in Section 2.2.2 for the sake of completeness.

2.2.1 Statistical Approaches

2.2.1.1 Term Weighting in Information Retrieval

The earliest term weighting method for Information Retrieval can be dated back to the work of Luhn [58]. The basic idea is that each term in a text collection can be classified into one of three frequency categories: high, medium and low frequency. Terms with medium frequencies are the best for indexing and searching, while the low and high frequency terms have minimal and negative impact on retrieval, respectively.

A more elaborated weighting method based on the frequency of occurrence is TF-IDF [73], where the weight of a term i in a document j (w_{ij}) is determined by the frequency of the term i in the document j (tf_{ij}) and the number of documents containing term i in the text collection (df_i).

$$w_{ij} = tf_{ij} \cdot \log \left(\frac{N}{df_i} \right)$$

where N is the total number of documents in the text collection.

Salton et. al. [75] proposed a non-frequency based method, where a term is weighted by its “Discrimination Value” (DV) among documents.

Let t be a term, then the DV of t , i.e. $DV(t)$, is defined as:

$$DV(t) = ADS - ADS_t$$

where ADS is the average inter-document similarity without t , which can be computed by some appropriate similarity functions, and ADS_t is the average inter-document similarity with t . If t is capable of discriminating different documents, the value of ADS_t tends to be small – because the similarity among documents decreases when t is included, and the value of $DV(t)$ tends to be large.

The problem of the traditional weighting approaches applied in Information Retrieval is that the importance of a term is solely judged by its ability to distinguish different documents in a text collection for the retrieval process, not by its topicality in a domain. The term “astronomy” for example, which is clearly topical in the astronomy domain, will receive a rather low weight for a retrieval task on the basis of an astronomical document collection, because as a general term, “astronomy” is likely to be contained in many documents in the collection and is therefore not capable of distinguishing between these documents.

2.2.1.2 Term Weighting in Text Classification

For the task of text classification, a series of term weighting approaches has been developed in machine learning to reduce the high dimensionality of the feature space in text corpora. Such methods are usually referred to as the feature selection methods.

The most commonly applied feature selection methods include Information Gain (IG), Cross Entropy (CE), Mutual Information (MI), and Odds Ratio (OR) [88][62].

Let $\{c_i \mid i=1, \dots, m\}$ denote the set of categories in the target space. The weights of a term t according to the respective features selection method are calculated as below:

$$IG(t) = P(t) \sum_{i=1}^m P(C_i | t) \log \left(\frac{P(C_i | t)}{P(C_i)} \right) + P(\neg t) \sum_{i=1}^m P(C_i | \neg t) \log \left(\frac{P(C_i | \neg t)}{P(C_i)} \right)$$

$$CE(t) = P(t) \sum_{i=1}^m P(C_i | t) \log \left(\frac{P(C_i | t)}{P(C_i)} \right)$$

$$MI(t) = \sum_{i=1}^m P(C_i) \log \left(\frac{P(C_i, t)}{P(C_i) \cdot P(t)} \right)$$

$$OR(t) = \log \left(\frac{P(t | pos) \cdot (1 - P(t | neg))}{(1 - P(t | pos)) \cdot P(t | neg)} \right)$$

In these formulae, probabilities can be interpreted on an event space of documents, if the frequency of t in a document is simplified to a binary value of 1 or 0, representing the presence or absence of t in a document. $P(t)$ then indicates the probability that t occurs in a random document, $P(\neg t)$ the probability that t does not occur in a random document, and $p(C_i, t)$ the probability that t occurs in a random document and this document belongs to category C_i .

In the last formula, $P(t|pos)$ is the conditional probability of term t occurring given a positive class – a collection of documents representing the target domain and $P(t|neg)$ is the conditional probability of t occurring, given a negative class – a collection of documents not representing the target domain.

Some comparative studies were carried out to evaluate the performance of different feature selection methods for text categorization [62][88]. It has been shown that Odds Ratio performs significantly better and is more stable than other measures. Cross Entropy provides the second best performance, while Mutual Information performs slightly better than a “random” approach, where no feature selection method was applied, and Information Gain is worse than random. The reasons for the poor performance of the last two measures lie in the fact that the Mutual Information is not correctly normalized, and thus strongly influenced by the marginal probabilities of terms [88]; Information Gain also assigns high weights to features that are characteristic for negative class and tends to give contrary results as Odds Ratio [62].

Although feature selection in text classification seems to be similar to concept extraction at first glance, they are quite different in nature: while in concept extraction a term is weighted according to its ability to indicate the topic of a target domain, feature selection approaches weight a term according to its ability to assign documents to proper categories. Therefore, although many feature selection methods also consider different pre-defined classes of documents, some of them (such as Information Gain and Cross Entropy) compute term weights with respect to the whole set of classes, not to an individual class. Other feature selection methods such as Odds Ratio treat a multi-class classification task as multiple binary classification tasks, and are thus able to better weight a term with respect to a specific class by distinguishing the distribution of a term in the positive document set (i.e. documents contained in the target class) with its distribution in the negative document set (i.e. documents contained in other classes). This principle resembles that of the F_dF_c approach for concept extraction, which will be described in more detail in Chapter 3. It is worth noting that both Odds Ratio and F_dF_c tend to give high weights to low frequency terms occurring only in a target class. Such terms usually play an important role in text classification, but are too trivial to be

extracted as domain concepts. A detailed comparison between the Odds Ratio and other approaches to concept extraction can be found in Chapter 3.

2.2.1.3 Term Weighting in Concept Extraction

In the task of concept extraction for a domain, a term is weighted according to its ability to indicate the topic of the domain.

Liu et. al. [56] combined a simple statistical weighting method (term frequency) with heuristic rules to mine topic-specific concepts from the top ranked result pages of a web search engine by using the name of the topic as search query. One of the heuristic is for example to extract all contents between the emphasizing html tags like `<h1>`, `<h2>`, ..., `<h4>`, ``, `<i>`, `<u>`, etc., if the contents satisfy following rules:

- #Not containing person titles

- #Not containing URL or Email address

- #Not containing terms related to publication (such as conference, proceedings, etc.)

- #Not containing digits

- #Not too lengthy

In the extracted texts, stopwords are first recognized, and the text pieces between the stopwords are then extracted as itemsets. Those frequently occurring itemsets with the length greater than 3 will be accepted as concepts.

Some other weighting approaches like F_d/F_c [76] [38][87] and Kullback-Leibler divergence (KL) [22] are more closely related to our work. We will introduce them in more detail in Chapter 3 for a convenient comparison with our work.

2.2.2 Linguistic Approaches

Linguistic methods calculate the importance of a term mainly according to its lexical properties, e.g. the length of the term, the frequency that the term is contained in other

terms or the number of its modifiers. Such methods usually do not give domain-specific weighting.

Caraballo et. al. [10] gave weights to nouns based on the distribution of their modifiers. They first recognized nouns and their modifiers in a text corpus using appropriate linguistic softwares, then calculated the entropy of the rightmost pre-nominal modifier for each noun. General nouns tend to have high entropy values, since they usually have a complex modifier distribution; Specific nouns usually have fewer modifiers, and thus tend to have smaller entropy values.

Anick et. al. [1] considered two kinds of evidence for term weighting: “term dispersion” and “term spread”. The dispersion of a term is calculated as the number of modifiers of the term, which is similar to the principle of Caraballo, while the spread of a term is calculated as the number of documents containing the term. All terms with their dispersion greater than a threshold will be ranked by their spread, and the remainder will be ranked by their dispersion.

Another linguistic method is cost criteria, which is proposed by Kita et. al. [48].

$$K(a) = (|a| - 1) \cdot (f(a) - f(b))$$

where a is the target term that should be weighted (example: “contact lenses”) and b is every term containing a (example: “soft contact lenses” or “hard contact lenses”). $|a|$ is the length of a , i.e. the number of characters contained in a . $f(a)$, $f(b)$ are frequencies of term a and b respectively. In this weighting scheme, the length and the frequency of the target term make positive contribution to the final weight whereas the frequency of terms containing the target term makes a negative contribution. The basic idea hereby is that an important term should appear a significant amount of times by itself, but seldom appear in other terms. The problem arises with those general, but important terms in a domain, which usually occur quite often in other longer terms (such as the term “astronomy” in

the astronomy domain). These terms tend to have rather large values of $f(b)$ and their final weights are unfairly low.

Frantzi et. al. [33] improved the approach of Kita et. al. in that a fourth factor was added into the formula of cost criteria, i.e. the number of unique terms containing the term a – denoted as $c(a)$, so that Kita’s weighting scheme is modified as

$$\text{C-value} = (|a| - 1) \cdot \left(f(a) - \frac{f(b)}{c(a)} \right)$$

In this way, general terms that are distributed widely in different unique longer terms will be weighted more highly.

2.3 RELATIONSHIP DETERMINATION

2.3.1 Statistical Approaches

Semantic relationships between two concepts are usually based on common properties. Statistical relationship determination between terms follows a similar principle, using the context within which terms occur in a text corpus as properties.

Various notions of context and commonality of contexts are suggested in the literature: In some traditional approaches [73][25][13], the context of terms is defined as the text segments within which they occur individually, and common context is defined as the text segments within which they co-occur, where a text segment may be a document or a part of a document, e.g. a paragraph, a sentence, or a window surrounding the term with a certain size. These approaches are often referred to as co-occurrence analysis, where only the number of text segments is interesting, and the content of text segments, i.e. the terms contained in the text segments, is not taken into account. In contrast, other approaches [28][36][80] consider the content of text segments around individual terms as context, and define common context as the intersection of these

contexts. In the remainder we will refer to the first kind of context as *occurrence context* and to the second kind as *content context*.

These contexts and common contexts are usually quantified by means of similarity measures, which give similarity weights to each pair of terms to determine their relationships.

There exist numerous similarity measures which can be applied on occurrence context and/or content context, such as Cosine coefficient (COS), Dice coefficient (DICE), Jaccard coefficient (JAC) [70], pointwise mutual information (PWI) [18], χ^2 -test (CHI), Yule's coefficient of colligation Y (YY) [29], and the distributional similarity measures like L1 Norm, Contextual Jensen-Shannon Divergence [28] etc. We will introduce these measures in more detail in Chapter 4 to enable a convenient comparison with our work.

2.3.2 Linguistic Approaches

Linguistic approaches consider linguistic properties of terms in text corpora for determining their relationships.

Maedche et. al. [59] applied the Levenshtein Distance (LD) to measure the lexical similarity between two words. Given two words w_1 and w_2 , the distance is the number of deletions, insertions, or substitutions required to transform w_1 into w_2 . The greater the Levenshtein Distance, the more different the two words are expected to be. The problem of this method is that it only calculates lexical similarity, which does not necessarily reflect the semantic similarity between two terms. The terms “power” and “tower”, for example, have a very close Levenshtein Distance although they are semantically rather remote.

Other approaches used common grammatical content context of two terms to determine their relationships.

Grefenstette [39] recognized different kinds of modifiers of a noun, such as ADJ, NN (the noun is modified by an adjective or another noun), NNPREP (the noun is modified by a noun via a preposition), SUBJ, DOBJ, IOBJ (the noun appears as the subject, direct or indirect object of a certain verb). Each modifier is called an attribute of the noun. If two nouns have enough common attributes in a text corpus, they are regarded as to be closely related.

Numerous papers such as [72][44][60][54][11] used the same principle as Grefenstette for term relationship determination, although the grammatical attributes considered by different approaches may vary.

It is worth noting that the grammatical content context is a special type of content context described in the last section, whose recognition requires profound linguistic knowledge such as part of speech parsing, and well performed NLP softwares.

Other works such as [43] tried to find hyponyms from large text corpora by looking for some manually identified patterns like:

- #NP such as NP1, NP2, NP3...
- #such NP as NP1, NP2...
- #NP, NP1, or other NP
- #NP, NP1, and other NP
- #NP including NP and/or NP
- #NP especially NP and/or NP

2.3.3 Non-corpus-based Approaches

Besides the corpus based approaches introduced above, there also exist other alternative ways for term relationship determination.

Approaches like the one suggested in [59] take advantage of the structure of existing thesauri or ontologies. The relationship between two concepts can be determined by their common hypernyms, hyponyms or synonyms. The distance between different concept

locations in a thesaurus/ontology can also be used to indicate the relationships between concepts.

Chen et. al. [16] built web thesauri from web link structure with each web page being represented by a concept. After removing navigational links, relationships between web pages are discovered by using the following set of rules:

- # A link in a content page conveys an association relationship.
- # A link in an index page conveys an aggregation relationship.
- # If two pages have an aggregation relationships to each other, the relationship is changed to association.

As the next step, each web page will be represented by a concept induced from the anchor text over hyperlinks pointing to this web page. The relationships between web pages can then be applied to determine the relationships between concepts.

2.4 EVALUATION METHODS

Evaluation is one of the most difficult tasks in the process of automatic construction of domain-specific concept structures, since it is usually difficult to find universal criteria for effectively judging the relevance of a term or a relationship in a domain. The existing evaluation methods can be roughly divided into two groups: manual evaluation and automatic evaluation.

2.4.1 Manual Evaluation

In manual evaluation, test persons are usually engaged to check the quality of concept structures.

In a task of evaluating hierarchical topical structure [76], a user study was conducted with a group of eight users, who were asked to manually judge the “interestingness” of the extracted term relationships and classify the interesting relationships into four relationship types in WordNet.

In a work of automatic summarization of documents [50], a user study was carried out to investigate the advantage of using a term hierarchy for document summarization instead of using a ranked term list.

In an early work of Chen et. al. [13], a group of test persons were required to manually judge the quality of the thesauri that had been automatically constructed from several domain-specific databases. It was also investigated how well the thesauri could help the users to browse the databases.

Although regarded as more precise and meaningful than automatic evaluation, manual evaluation suffers from high experimental expenses, which is especially problematic for corpus based text processing tasks where various approaches with different parameters ought to be evaluated on the basis of a huge amount of data. Automatic evaluation is therefore preferable in our work because of its low cost and its high repeatability.

2.4.2 Automatic Evaluation

2.4.2.1 Evaluation in Applications

Instead of directly evaluating concept structures, some automatic evaluation methods integrate them in applications – usually Information Retrieval tasks – to see how well they can help to improve the overall performance of the system.

The earliest evaluation of this type can be dated back to the work of Salton [73] and Crouch et. al. [25], who used automatically constructed thesauri for query expansion in Information Retrieval tasks. An improvement of recall with 10-20% was demonstrated when the thesauri were applied in a similar environment as the one from which the thesauri had been originally derived.

A similar principle of automatic evaluation is adopted by many other works as shown in [46][16][60][1][30].

2.4.2.2 Using statistical measures

Lawrie [50] evaluated a concept structure, or more specifically, a hierarchy of terms in a document summarization task by computing the Expected Mutual Information (EMI) between the hierarchy and the original set of documents the hierarchy summarizes.

$$I(T, V) = \sum_{t \in T, v \in V} P(t, v) \cdot \log \frac{P(t, v)}{P(t) \cdot P(v)}$$

where T is the set of topic terms in the hierarchy and V is the set of non-stopwords occurring at least twice in the document set. The joint probability $P(t, v)$ is calculated as:

$$P(t, v) = \sum_{d \in D} P(d) \cdot P(t | d) \cdot P(v | d)$$

where D is the set of documents and $P(d)$ is a uniform distribution.

The EMI measures the extent to which the distributions of the topic terms and the vocabulary terms deviate from stochastic independence. The greater the dependence between the two random variables, the better the hierarchy summarizes the document set.

2.4.2.3 Comparing with Gold standards

Assuming that we could find a so-called gold standard concept structure, i.e. an ideal concept structure containing all interesting concepts and relationships in a domain, the automatic evaluation task would become fairly easy – we only need to compare our concept structure with the gold standard, and judge the quality of our concept structure by computing the overlap between the two.

In practice, however, such gold standards never exist. Some works tried therefore to manually build a concept set or a concept structure, and use it as a gold standard for evaluating automatic approaches. In the work of [72], for example, 59 words were randomly selected from 8257 extracted words. A human expert was required to construct a set of synonyms for each word, which were then used as a gold standard. The same

process, i.e. finding synonyms for the selected words, was repeated by using automatic methods adopting different similarity measures. The approach producing the results closest to the gold standard was regarded as the best measure. In an evaluation in [83], a student was asked to manually extract a list of important terms from a set of personal E-mails. This term list was then used as a gold standard to evaluate different concept extraction methods.

Other works used more extensive, pre-existing thesauri or ontologies as gold standard, whose construction usually requires long time cooperation of a (large) group of human experts. Lin [54] compared his automatically built word sense groups with Wordnet and Roget's thesaurus. In a task of automatic construction of taxonomy in the domain tourism, Cimiano et. al. [21] compared the automatically constructed taxonomy with a tourism ontology. They argued that the more the taxonomy resembles the ontology, the better the taxonomy will be regarded.

CHAPTER 3 AUTOMATIC EXTRACTION OF DOMAIN-TOPICAL CONCEPTS

In this chapter, we address the problem of how to effectively extract important terms as concepts in a domain. For the sake of convenience, we will henceforth refer to the domain for which concept structures should be built as the *target domain*, and the terms, whose importance are being calculated, as the *target terms*. As discussed in Chapter 1, there are two problems to solve for the task of concept extraction: 1. Choosing a suitable data source to represent the target domain. 2. Finding an effective method for extracting topical concepts from the data source. In Section 3.1, we discuss the possibility of using web directories as a data source. In Section 3.2, we present a new approach to weight term topicality and analytically discuss why this approach outperforms existing weighting methods. We then develop an automatic evaluation method and report on a series of experiments to validate our analytical assessments of different term weighting approaches in Section 3.3. In Section 3.4 we complement the automatic evaluation with a manual evaluation on a smaller sample of highly weighted terms. Some examples are provided for better illustration. Finally, we draw conclusions in Section 3.5.

3.1 WEB DIRECTORIES AS A SOURCE FOR COLLECTING DOMAIN-TOPICAL CONCEPTS

Choosing a suitable source to represent a target domain is the first important step in the whole process of construction of concept structures. In our view, the quality of such sources depends on the following aspects:

- Coverage: A good source should cover as much important information of a domain as possible.

- Accuracy: An ideal source should contain as little noise as possible, where noise refers to irrelevant information for a target domain.
- Up-to-dateness: A good source should be flexible enough to reflect not only stable domain knowledge, but also the most up-to-date information in a domain
- Accessibility: The source should be easy to access for further automatic processing.

According to these criteria, web directories like the Yahoo! directory, the ODP (Open Directory Project) and the Google directory (An extended version of ODP) are a suitable source to fulfill our task.

Web directories organize information in categories that are maintained by numerous human editors (65264 Editors with ODP, Oct. 2004). The editors, usually highly qualified domain experts representing the interests of different interest groups and diverse possible applications, add relevant web sites to the categories daily, so that the information in various domains can be enriched on a regular basis. The web sites themselves are maintained by countless authors. They are usually updated more frequently than web directories, and are capable of providing much more domain information. This property of web directories tends to provide very high flexibility and information coverage, which can hardly be reached by traditional newspaper archives or domain databases maintained by several editors.

With respect to accessibility, web directories are almost always accessible through the Internet – a compressed free copy of the complete content of ODP is regularly available for download on the ODP web site, while many traditional domain databases have restrictions on user access. It is worth mentioning that, besides web directories, web search engines like Google and Altavista could also be used to provide domain information, as shown in [16][56][82][4]. However, two drawbacks of search engines make them less appropriate than web directories for our task: first, it is generally difficult to choose the right keywords to optimally represent a domain, which is crucial for retrieving high quality search results as data sources; second, most of the search engines

have access restriction on search results for normal users, as in the case of Google, which restricts its results to around 80 pages with 10 hits per page.

The “weak point” of the information in web directories, similar to other information on the WWW, is its relatively low quality. Human maintainers of web categories only consider whether a web site as a whole is suitable for a domain. As a web site inevitably contains noise, i.e. irrelevant information for a domain, this noise is also included in the data source, like the terms “click me” and “contact”. Such terms often occur on both domain-specific and domain-unspecific web pages, and are therefore likely to be included in the data source together with the domain-specific web pages carrying them. It is crucial to develop efficient weighting schemes to filter this noise and assign better weights to important concepts.

3.2 TOPICALITY BASED TERM WEIGHTING

3.2.1 Computing Term Topicality

The goal of term weighting is to provide a quantitative measure for computing the importance of terms. The definition of term importance may vary in different applications. In traditional Information Retrieval, the importance of a term is judged by its ability to distinguish different documents, such as the TF-IDF weighting method. In our work, however, the importance of a term depends much more on its ability to indicate the topic of a target domain, i.e. its topicality in the target domain.

We believe that the topicality of a term is a combination of two factors: *term representativeness* and *term specificity*. While term representativeness measures how well a term is capable of representing a domain, term specificity indicates how specific a term is in the target domain rather than in the whole text collection, which may contain various other domains besides the target domain.

Let us imagine a situation in which one is required to manually build a concept structure for a specific domain, say astronomy. As described in Chapter 1, the first step for achieving this task is to select a certain number of the most important concepts in the domain. Assuming the number is restricted to 5, one reasonable answer could be “astronomy”, “star”, “galaxy”, “solar system” and “telescope”. Obviously, such terms are specific to the topic astronomy and are unlikely to be equally distributed in other topic domains. On the other hand, as the top 5 most important terms in the domain astronomy, they should be capable of covering the whole topic area as widely as possible, that is, they have to be fairly representative of the target domain. In practice, although the number of concepts in a domain-specific concept structure is usually much larger than 5 (usually ranging between thousands and tens of thousands), the same principle for concept extraction still holds, that is, the topical terms should be both representative and specific for the target domain to be extracted from the numerous terms in a data source.

As observed in Chapter 2, there exists a series of term weighting approaches that have been applied for similar tasks. The difference between these methods lies in two points: 1. Whether the representativeness aspect is taken into consideration, as many approaches only compute term specificity for concept extraction. 2. How the term specificity is calculated. In the next section, we present a detailed analysis of several most relevant approaches for concept extraction with respect to these two points.

3.2.2 $F_d F_c$, Odds-Ratio and Kullback-Leibler Divergence

In this section, we will review three most important state-of-the-art weighting methods, including $F_d F_c$, Odds Ratio and Kullback-Leibler Divergence (KL), with respect to the different ways they calculate term importance.

$F_d F_c$ is the most commonly applied weighting method for concept extraction [76][38][87], usually computed as:

$$F_d - F_c(t) = \frac{f_d(t)}{f_c(t)}$$

where $f_d(t)$ denotes the frequency of a term t in a target domain d and $f_c(t)$ the frequency of t in the whole text collection c . The basic idea of the $F_d - F_c$ method agrees with our intuition for computing term specificity, that is, the more a term appears in a target domain and the less it appears in the whole text collection, the more it will be regarded as specific to this domain, and thus the higher it will be weighted. Since term representativeness is not considered in $F_d - F_c$, a term is actually weighted only according to its specificity, not its topicality, which may lead to an unsatisfactory performance in concept extraction, as will be shown in the next section. As there is still no official name for this measure to the best of our knowledge, we will henceforth refer to it in our work as the $F_d - F_c$ measure.

As a weighting method originally developed for text classification, Odds Ratio [62] applies a term weighting principle similar to the $F_d - F_c$ measure, and thus can be also used for concept extraction.

$$Odds\ Ratio(t) = \log\left(\frac{P(t|pos) \cdot (1 - P(t|neg))}{P(t|neg) \cdot (1 - P(t|pos))}\right)$$

where $P(t|pos)$ is the conditional probability of term t occurring given a positive class – a collection of documents representing the target domain and $P(t|neg)$ is the conditional probability of t occurring, given a negative class – a collection of documents not representing the target domain. Similar to $F_d - F_c$, Odds Ratio also only calculates term specificity, not term topicality. The difference between the two measures lies in the fact that, instead of using the whole text collection for comparison as in $F_d - F_c$, Odds Ratio uses the negative document set as the comparison basis. Notice that the log function in Odds Ratio increases strictly monotonically, and thus does not have any effect on ranking terms according to their weights.

The Kullback-Leibler divergence (KL) [22] was originally used for comparing two probability distributions. Lawrie [50] and Cronen-Townsend et al. [23] firstly employed this measure to calculate term topicality.

$$KL(t) = P_d(t) \cdot \log\left(\frac{P_d(t)}{P_g(t)}\right)$$

where $P_d(t)$ is the probability that a term t appears in a domain d and $P_g(t)$ the probability that t appears in general English. $P_d(t)$ can be estimated by dividing the frequency of t in the domain by the total frequency of all terms in the domain. $P_g(t)$ can be estimated in the same way. General English can be simulated by the whole text collection containing the domain.

The KL measure applies the factor $P_d(t)/P_g(t)$ to calculate term specificity, which is in principal similar to F_d/F_c , – i.e. comparing the distribution of a term in a domain with that in the whole collection. Their difference lies in two aspects. First, KL is able to compute term representativeness by multiplying the term specificity factor with another factor, $P_d(t)$, which is a reasonably good measure for calculating how representative a term is in a domain. Second, $P_d(t)/P_g(t)$ is put into a log function to bias those unspecific terms more strongly, that is, if $P_d(t)/P_g(t)$ is smaller than 1, $\log(P_d(t)/P_g(t))$ will be negative. A noisy term, for example, is likely to be widely distributed in a target domain, and will thus have a large value of $P_d(t)$. When directly combined with the factor $P_d(t)/P_g(t)$ without using log function, the term will still have a chance to be weighted higher than a topical concept because of the large value of $P_d(t)$, although the value of $P_d(t)/P_g(t)$ may be rather small, say, smaller than 1. However, when the log function is used, $\log(P_d(t)/P_g(t))$ will become negative, and the value of KL measure, which is a multiplicative combination of $P_d(t)$ and $\log(P_d(t)/P_g(t))$, will also be negative. In this case, a larger value of $P_d(t)$ only leads to a smaller value of KL.

Although the approaches introduced in this section may improve the traditional weighting methods to some extent, we still find some of their principles unsatisfactory. In

the next section we propose a novel approach for concept extraction – the distributional topicality weighting, which considers both specificity and representativeness for computing term topicality, and is able to perform a better specificity calculation by comparing term distribution in a target domain with that in different reference domains.

3.2.3 Distributional Topicality Weighting

3.2.3.1 Topicality vs. Specificity

As shown in the previous section, some weighting approaches like F_d/F_c and Odds Ratio solely measure specificity instead of topicality when giving weight to a term. Although some of them prove to work well in the task of feature selection for text classification [62] [88], there is no evidence that they are also suitable for the task of concept extraction in a target domain. In contrast, we expect a rather bad performance when directly applying them for this purpose. The reason is, as indicated in previous sections, that term specificity only builds one aspect of term topicality. Using it alone for term weighting leads to the problem that many trivial terms appearing only in the target domain will receive the highest weights, and will be incorrectly ranked higher than those really topical terms, which may occasionally appear in other domains.

In the task of text classification, it might be sufficient to consider only term specificity for feature selection, because the trivial but special terms usually play a key role in assigning a target document to the correct category. For extraction of topical concepts from a target domain, however, such terms should not be ranked at the top of the candidate term list due to the lack of representativeness calculation. A human expert building an astronomical concept structure, for example, will not choose terms like “Birmingham Astronomical Society”, “Federal Star Registration”, “Darian Defrost Calendar” as the most important concepts, although they are quite specific to the astronomy domain. Instead, he will rather prefer terms like “astronomy”, “telescope”,

“planet”, “X Ray” etc., which are both representative and specific, i.e. topical, in the target domain.

Representativeness weighting also plays an important role for hierarchical relationship determination among extracted concepts, which will be addressed in detail in the next chapter. The intuition behind many approaches to hierarchical relationship determination [50][76] is that a “broader” term is more representative than a “narrower” one. It is interesting to note that in the work of [76], from which the $F_d F_c$ measure originated, a representativeness based term weighting is implicitly employed when calculating the so-called subsumption relationship – a hierarchical relationship that is automatically determined by comparing document frequency of terms. Based on this relationship, a concept hierarchy can be automatically constructed, with high representative terms at the top and unrepresentative terms at the bottom of the hierarchy. As a concept hierarchy is usually browsed in a top-down manner, the terms at higher levels are usually regarded as more important than the terms on lower levels, which is principally similar to a representativeness based term weighting.

Normally, representativeness can be successfully calculated using rather simple measures. In KL-Divergence for example, the representativeness of a term is computed as its relative term frequency in the target domain. This simple principle of representativeness calculation applies well in our work, because web documents in the appropriate web directories usually provide sufficient domain coverage, which guarantees a wide distribution of highly representative terms in the data basis. With the noisy terms being filtered out by a well-developed specificity measure, the term frequency or the document frequency of a term is a reasonably good indicator for the representativeness of the term.

3.2.3.2 Improvement of specificity calculation

In contrast to representativeness, specificity calculation is more complex and the existing approaches can still be improved. It is clear that term specificity can only be accurately calculated by comparison. Previous approaches only use the whole text collection or the negative text collection as a basis for comparison, without distinguishing the individual domains contained in the text collections. This can lead to unsatisfactory weighting.

As an example, let us assume that there are a total of 6 different domains existing in the whole text collection. We also make a trivial assumption that these 6 domains, or rather, the text sets representing the domains, are of similar size, so that we can use the absolute term frequency in the following instead of the relative term frequency for a more intuitive depiction. As shown in Figure 3.1, Term 1 is evenly distributed in all domains with the same frequency 10; Term 2 appears only in Domain 2 and Domain 5, with a frequency of 20 and 100 respectively. Obviously, Term 1 is not specific for any domain (such as the term “click me” – a rather general term that could be found on almost every web page). Term 2 is specific for both Domain 2 and Domain 5, although its frequency in Domain 2 is much smaller than that in Domain 5 (Example: the term “X Ray”, which is specific both to the astronomy and physics domains, although it may have relatively lower frequency in astronomy as compared to physics).

If we use F_d/F_c to calculate specificity (KL and Odds Ratio follow a similar principle for computing term specificity), Term 1 and Term 2 will have a same specificity value of $1/6$ for Domain 2 ($10/60$ with Term 1 and $20/120$ with Term 2), meaning that the two terms are equally important to this domain, which is obviously wrong. The problem lies in the fact that many terms could be specific to more than one domain. Previous approaches simply compare the target domain with the whole/negative text collection,

without distinguishing other individual domains contained in the text collection. They thus fail to calculate the specificity of multi-specific terms.

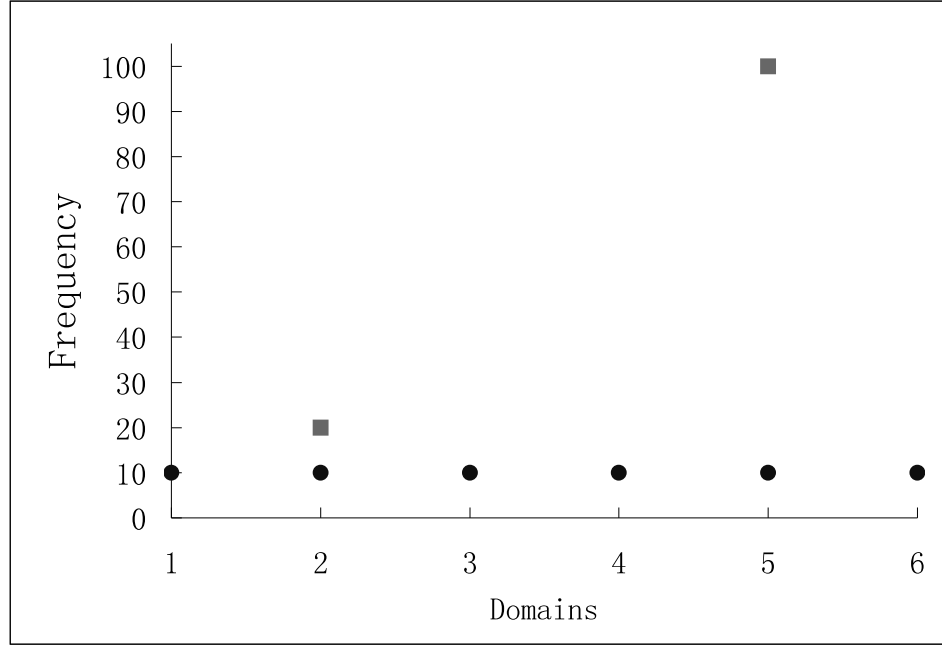


Figure 3.1. Domain distribution of Term 1 and Term 2

For a better calculation of term specificity, we distinguish different domains in the text collection rather than using the whole text collection or the negative text collection. Intuitively, if a term is evenly distributed among all individual domains, it should have a low weight of specificity. The distribution grade of a term can be readily calculated on the basis of the variation coefficient, which normalizes the variance of a distribution by its mean:

Given a term t , $f_i(t)$ is the frequency of t in Domain i – it could be either absolute frequency if the domains, including the target domain and the reference domains, are of similar size, or relative frequency if the sizes of domains vary to a great extent. The total number of domains is m .

$$\text{Mean Value: } MV(t) = \frac{1}{m} \cdot \sum_{i=1}^m f_i(t)$$

$$\text{Distribution Grade } DG(t) = \frac{1}{MV(t)} \cdot \sqrt{\frac{1}{m} \cdot \sum_{i=1}^m (f_i(t) - MV(t))^2}$$

The distribution grade measures the variance of the frequencies, with which the term appears in different domains. The smaller the value, the more evenly the term is distributed, and the more unspecific it is. For the example above, the DG value of Term 1 is 0, while the DG value of Term 2 is 1.826, which is much larger than that of Term 1, indicating that Term 2 is not evenly distributed in the text collection, and is specific to the domains in which it occurs.

As in the KL measure, we can also apply a log function to DG to bias the unspecific terms more strongly. The factor for computing term specificity will then be changed to $\log(DG)$.

Since this method considers the distribution of a term in different individual domains for computing its specificity in a target domain, we refer to it as the distributional specificity weighting.

3.2.3.3 Computing Term Topicality

Finally, we need to combine the weighting measures developed in previous sections to compute term topicality. Because a term needs to be both representative AND specific in order to be topical, topicality is clearly a conjunctive combination of representativeness and specificity. Since the usual way to implement a conjunctive combination of different factors in Information Retrieval and Concept Extraction is to multiply the factors, like the TF-IDF measure [73] and the KL measure [50], we also multiply specificity and representativeness here to calculate topicality.

$$TW(t) = doc_num(t) \cdot DG(t)$$

where $TW(t)$ is the topicality weight of a term t , $doc_num(t)$ denotes the document frequency of t and $DG(t)$ is the distribution grade of the term.

Notice that we choose document frequency here to calculate term representativeness, which is the number of documents in which the term occurs. Our experiments show that this yields better results than using the absolute term frequency. The reason for this is that a term distributed widely in a domain with relatively smaller term frequency in individual documents will represent this domain better than a term which appears with large frequency only in a small number of documents.

A modified version biasing unspecific terms more strongly is proposed as:

$$TW(t) = doc_num(t) \cdot \log(DG(t))$$

This measure particularly deals with those unspecific terms such as “click me” or “welcome” which often have fairly large *doc_num* values, and quite small *DG* values (usually smaller than 1). By adding the log function, the specificity weight, i.e. $\log(DG)$ will become negative, and the large value of *doc_num* will only result in an even smaller negative value of *TW*. This is in principle similar to the KL measure which also applies a log function to bias unspecific terms more strongly.

3.3 AUTOMATIC EVALUATION

In this section we intent to automatically evaluate the quality of a web category in representing a target domain, and the performance of different weighting methods for concept extraction.

Experiments are carried out for two target domains: Astronomy and Construction (including both Civil Engineering and Architecture), respectively. For each target domain, we employ a manually-built and well accepted concept structure as a gold standard, which are expected to contain the most important concepts and relationships in a domain, and then compare the experimental results with the gold standards.

For evaluating the quality of web categories, we extract all terms from web categories without weighting them, and compare the terms with the concepts contained in

the gold standards. We hypothesize that web categories are capable of covering a large part of gold standards, and thus can provide sufficient information to represent target domains.

For evaluating weighting approaches, we use different weighting methods to weight and rank candidate terms. We assume that a better weighting method will supply more concepts within a certain number of top-ranked candidate terms, and thus will have a larger overlap with the gold standards within this threshold. We hypothesize that the two topicality based weighting methods of us will outperform other concept extraction approaches for a wide range of candidate terms on different kinds of data basis.

3.3.1 Principle of Automatic Evaluation

3.3.1.1 Comparing with Gold Standard

Evaluation is known as one of the most difficult tasks in Information Extraction and Information Retrieval research. A complete manual evaluation is hardly feasible for the problem at hand because of the large number of candidate terms (normally 10,000 ~ 10 million depending on the size of the target domain), which would need to be individually evaluated with respect to their topicality. It is even more difficult if the evaluation needs to be done repeatedly for each different weighting approach and each different tuning alternative for comparison. Moreover, manual evaluation always brings problems with respect to intersubjective disagreement and assessment errors. An automatic evaluation method has to be developed.

There exist many so-called gold standard concept structures in different domains, which are developed manually by human experts with high domain expertise. They are usually widely accepted and are assumed to contain the most important concepts in a domain. A gold standard concept structure can be given in different forms, e.g. a thesaurus, an ontology or a domain lexicon.

For evaluating the quality of web directories as a suitable data source, we compare the terms selected from the relevant web directories with the items in a well-known gold standard. A large overlap indicates that web directories are capable of providing sufficient topical concepts in a target domain, and can therefore be used as a good source for our task.

Using gold standards to compare different weighting approaches is more difficult, because most of the gold standards do not assign weights to their member terms. In the following, we propose a new method to work around this difficulty.

Let us imagine a real life application of manual ontology construction. Human experts building an ontology hope to find topical terms from observed candidate terms. If the number of candidate terms is large, they usually sort these terms by applying a term weighting scheme, with important terms being ranked at the top. They then set a threshold to determine the number of candidate terms they want to further consider. A good weighting method tends to rank more topical terms above this threshold. If the experts change the threshold, the weighting method should perform in a stable way for the different threshold values.

Based on this intuition, we first sort the candidate terms in descending order according to their weights. It is clear that different weighting methods will result in different orders. We then set a threshold to limit the number of the top terms for comparison. The idea is that a good weighting method will supply more topical terms within these top terms than a bad one. (A term will be judged as being topical if it is contained in a gold standard). In the process of manual construction, this means that an ontology constructor will have chance to find more “interesting” terms among the terms he observes, when the terms are sorted by a good weighting method.

One problem of using gold standards for evaluation is that the terms not contained in gold standards may also be “interesting”, which are not taken into account in evaluation. However, it is necessary to point out that the goal of our evaluation is to compare

different weighting approaches, not to evaluate the absolute performance of the individual methods. If one weighting method performs better than another one when compared with “gold standards”, it is reasonable to assume that the same will hold for the whole set of “correct” terms, which are rarely fully retrieved in practice.

3.3.1.2 The F-Measure

Based on the automatic evaluation principle, recall and precision of a certain weighting method can be defined as follows to measure its ability to include gold standard terms (i.e. topical terms) within a threshold (i.e. the number of observed candidate terms):

$$\text{recall } R = \frac{y}{L} \quad \text{precision } P = \frac{y}{x}$$

where x is the number of observed candidate terms with their weights larger than the threshold; y is the number of gold standard terms contained in x ; L is the total number of terms in the gold standard.

If we further define the total number of candidate terms as M and the number of gold standard terms in M as n , there should be $0 < y \leq n \leq L$ and $0 < x \leq M$. In our experiment, M is usually much larger than L , i.e. $M \gg L$.

The F-measure [70] is applied for combining the effect of both Precision and Recall:

$$F = \frac{2 \cdot P \cdot R}{(P + R)}$$

Taking x as the x-axis and F as the y-axis, we can directly compare different F-measure curves for different weighting methods.

In Figure 3.2, each curve (excluding the origin (0, 0)) represents a weighting method. All curves converge at the point $(M, 2n/(M+n))$, where both candidate terms and topical terms are exhausted. Curves (2) and (3) are two special curves. If no weighting scheme is applied and the candidate terms are sorted in a random order, then the topical terms are

expected to be evenly scattered among all candidate terms. Its F-measure is shown as Curve (2). Curve (3) shows an ideal case of term weighting. Assuming an ideal weighting method is able to rank all topical terms on the top of the list, then the top n candidate terms are all terms in the gold standard. In this case, the curve will quickly reach the point $(n, 2n/(n+L))$ and then, since precision decreases rapidly and recall remains constant, the curve drops quickly until the point of convergence.

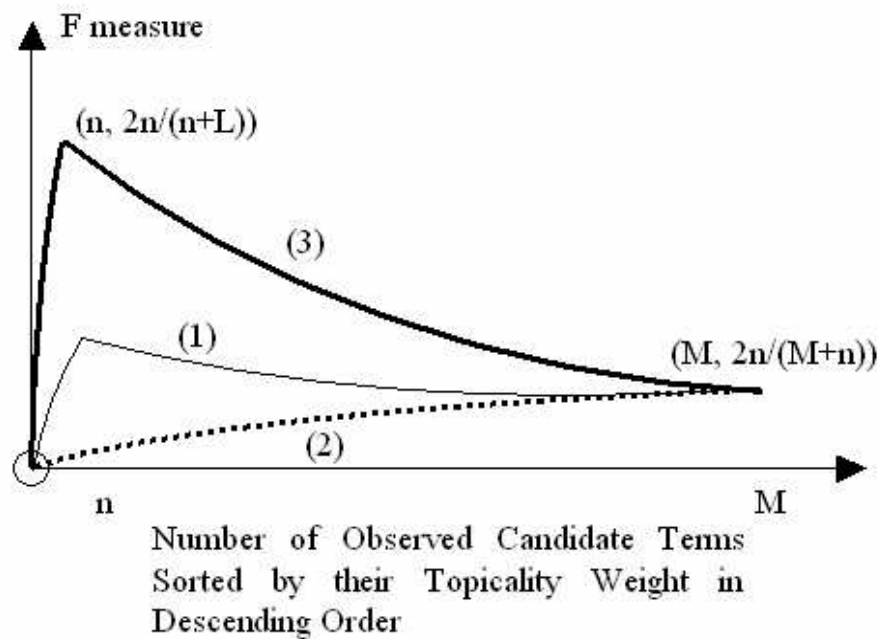


Figure 3.2. Theoretical F-measure curves of different weighting methods

All good curves, e.g. Curve (1) should lie between Curve (2) and Curve (3). Like the ideal curve, they will sooner or later reach a maximum point, then go down until the convergence point. The maximum points of all real curves will be located at a larger x -value than that of the ideal curve.

A curve lying above another curve at a certain x value means that the first weighting method supplies a greater F-measure, and thus is better than the second one at this x value.

Note again that in practice, the number of interesting terms in a domain usually greatly exceeds the number of concepts in a domain specific gold standard. Thus, when applied in a real life application, the maximum points of all curves in Figure 3.2 tend to shift greatly to the right. The best performance of each weighting approach will then be achieved with a much larger number of candidate terms.

3.3.2 Experimental Setup

In our experiment, we have chosen astronomy as the target domain for which the topical terms are selected. This has several advantages: first, the size of the domain is large enough to achieve a satisfactory evaluation, but not too large to bring unnecessary burden to the experiment; second, as a scientific domain, astronomy has been the subject of research for quite a long time, therefore the field is relatively stable; third, there exist several gold standard concept structures in the domain astronomy that can be used for comparison.

We have also chosen 17 reference domains for computing term specificity. 9 of them, together with the astronomy domain, are all in the same parent domain of science, including mathematics, physics, chemistry, agriculture, biology, ecology, energy, engineering and geography. The other 8 domains have less in common with the domain of science, including construction, finance, law, military, computer/networking, pharmacy, psychology and swimming.

As the gold standard we have taken a well-known astronomy thesaurus⁷, which is endorsed by the International Astronomical Union and has been compiled to standardize the terminology in the field of astronomy. We downloaded the whole thesaurus. All terms (around 2900) in it were parsed and saved in a database for further processing.

⁷<http://www.aao.gov.au/lib/thesaurus.html>. Compiled by Robyn M. Shobbrook (Anglo-Australian Observatory) & Robert R. Shobbrook (Sydney University, School of Physics)

We use both the Yahoo! directory and Google directory as the base web directories. In order to collect domain-specific data from the web directories, we first locate the appropriate categories in the web directories. We then crawl the URLs of all web sites under the categories and their subcategories. For simplicity, all cross-reference domains with the symbol @ in Yahoo! and all subdirectories pointing outside the Google directory “science/astronomy” are ignored. After deleting the overlap, the URLs of web sites taken from the Yahoo! and the Google directories are merged. For the domain of astronomy, we collect around 8000 web sites from Yahoo! and Google directories. The content of each given URL is also crawled. Since the emphasis of our work is not recognizing the important parts of a web site, we simply crawl the first page under each given URL, which results in around 8000 HTML pages with a total size of 128 megabytes.

Since many web pages have the meta-data “keywords” and “description”, containing abstract information about the web site, we are also interested in how the data from the web directories will overlap with the gold standard if only the information from “keywords” and “description” of a page is used. Therefore, two data bases are prepared in our experiments: one using only the “keywords” and “description” information and the other based on the whole home page of each web site.

In each data basis, we first delete all html tags. Stopwords are recognized by using a common stopwords list. The words between two stopwords are treated as a pseudo-phrase, if there are no other separators except space among them. These pseudo-phrases – we call them *word groups* – are used to do the first step comparison with the terms in the gold standard. As the second step, we do some further processing on these word groups. We choose only those word groups with frequency greater than 2, and perform POS (Part of Speech) tagging on this set of word groups by using the MontyTagger [57]. The word groups satisfying the regular expression {adj ? noun +} are used for the second step comparison with the gold standard. Since we are interested in “real terms” rather than the reduced stems of terms when building concept structures, no stemming is carried out in

our experiments. We also decide not to carry out lemmatization, because the inflected form of a term may sometimes have a completely different meaning as compared to the base form (Windows/Window, News/New). Such inflected forms may be more important than or as important as their base forms, and tend to occur more often in some domain-specific text collections. It is therefore desirable to also include these inflected forms in our concept structures.

When comparing with the gold standard, we only consider the so-called “Direct Match”. Two terms will only be accepted as a “Direct Match” if they are completely identical regardless of case or if the difference can be recognized by automatic methods using a few simple normalization rules, such as transforming “-“ to a whitespace, so that the candidate terms “x ray” and “x-ray” will both be directly matched to “X RAY” in the gold standard.

It is worth pointing out that many candidate terms that are not recognized as “Direct Match” may be synonyms of the terms in the gold standard. For example: “21 cm line” in candidate terms as opposed to “21 cm radiation” in the gold standard, and “Totality Zone” as opposed to “Zone of Totality”. Although such terms will be easily identified as a “Match” by a human expert, they cannot be recognized automatically, and are thus not considered in our evaluation.

3.3.3 Evaluating the Quality of Web Directories as a Source

By comparing the terms selected from the astronomy-relevant web categories with the items in the astronomy thesaurus, the quality of web directories as a source can be evaluated. Experimental results are shown in Table 3.1 and 3.2.

The results in the two tables indicate that by using only a small part of each web site contained in the web directories (first page or keywords/description on the first page), we are able to achieve a good overlap with a well accepted domain thesaurus. It is reasonable to assume that if we do further content analysis on the web sites as shown in the work of

[15], and extract more useful information from them, this overlap will continually increase. This confirms our assumption that web directories provide valuable domain information.

As web-based search engines like Google are becoming more and more important, they are also often used to build domain sources as shown in [16][56][82] where the domain name is submitted as a query to the search engines and the retrieved results serve as the source for extracting topical concepts. Here, the validity that a web site belongs to a domain is determined by the automatic algorithm of the search engines, while in web directories it is assured by human editors.

Table 3.1. Only using “keywords” and “description” – Web directories

	Number of Candidate Terms	Direct Match in Gold Standard
First Step Comparison	43740	953 (32.86%)
Second Step Comparison (with POS and Freq>2)	7139	538 (18.55%)

Table 3.2. Using the whole page – Web directories

	Number of Candidate Terms	Direct Match in Gold Standard
First Step Comparison	3650001	1746 (60.21%)
Second Step Comparison (with POS and Freq>2)	48986	1300 (44.83%)

In order to check the quality of search engine results as data sources, we have submitted the keyword “astronomy” to Google and crawled all web sites in the results (around 800 hits because of the restriction of Google). The information is processed in

the same way as for web directories. In Table 3.3, we show the results by using the whole home page of each web site.

Table 3.3 Using the whole page – Google

	Number of Candidate Terms	Direct Match in Gold Standard
First Step Comparison	58115	971 (33.48%)
Second Step Comparison (with POS and Freq>2)	7552	549 (18.93%)

Table 3.3 shows that Google is also capable of supplying good overlap with the gold standard. In fact, when compared with the same level of candidate terms, it supplies almost the same amount of “Direct Matches” as the web directories in Table 3.1. Another interesting point is that among the 800 astronomy-specific web sites retrieved from Google, only 236 web sites are also contained in the web directories. The common “Direct Match” number in the second step comparison in Table 3.1 and Table 3.3 (538 and 549 respectively) is only 399. However, the common “Direct Match” number in the second step comparison in Table 3.2 and Table 3.3 (1300 and 549 respectively) is 541, which is almost the total number of “Direct Match” in the second step comparison in Table 3.3.

The biggest problem of using search engines as a source is their restriction on the number of hits, so that a normal user cannot retrieve enough information for a domain by using search engines (around 800 web sites with Google). Web directories, especially ODP, do not have any access restriction, which means that a normal user can easily get much more relevant information about a domain (in our experiment, around 7000 web sites are collected for the astronomy domain from the web directories).

3.3.4 Evaluating Weighting Approaches for Concept Extraction

Since the second step comparison in Table 3.2 provides a relatively small number of candidate terms (48986) and a reasonably large overlap with the gold standard (45%), we decide to use it as the data basis for the comparison of different weighting approaches, which will be henceforth called the data basis 1.

For computing the distribution grade, i.e. DG in our methods, we also crawl all 17 reference domains from the Yahoo! and Google web directory, resulting in about 1.2 gigabytes of raw data. All data are processed in the same way as in the domain astronomy. These 17 reference domains form the “negative collection” for the computation of Odds Ratio and all 18 domains together, including the domain astronomy, build the “whole collection” for computing the $F_d F_c$ and the KL measure.

By applying the automatic evaluation method presented in Section 2.3, we are able to draw different curves for different weighting approaches in Figure 3.3, including TF-IDF, $F_d F_c$, Odds Ratio, KL and the two methods suggested by us: $\text{doc_num} * \text{DG}$ and $\text{doc_num} * \log(\text{DG})$. For the purpose of illustration, we also draw the curve for the random ranking without using any weighting method. As indicated in Figure 3.2, the ideal curve is much higher than all of the curves of the weighting methods with a maximum F value of 0.73, which lies much higher than other “real” curves. For a better scaling of the other curves, we do not draw the ideal curve on the figures.

It is clear that, except for a small part at the very beginning (within the first 2000 – 4000 candidate terms), our $\text{doc_num} * \text{DG}$ curve lies clearly above other curves in most parts of the diagram. The curve representing KL performs well at the very beginning, it rises together with our curve very quickly to a similar maximum value. However, it drops very quickly after the maximum point and even goes beneath the TF-IDF curve after 10000 terms. $\text{doc_num} * \log(\text{DG})$ behaves similarly as $\text{doc_num} * \text{DG}$, with most parts of its curve exceeding the competing methods. However, after the maximum point, it runs a little worse than $\text{doc_num} * \text{DG}$. In fact, it almost completely overlaps the TF-

IDF curve after 22000 observed candidate terms. At the very beginning, it seems to outperform the KL curve to a small extent. Odds Ratio and $F_d F_c$ perform clearly worse than other approaches. The reason is, as indicated in the previous sections, that these two measures only calculate term specificity, which is merely one aspect for computing term topicality, suggesting that purely specificity based weighting is not suitable for the concept extraction in our task.

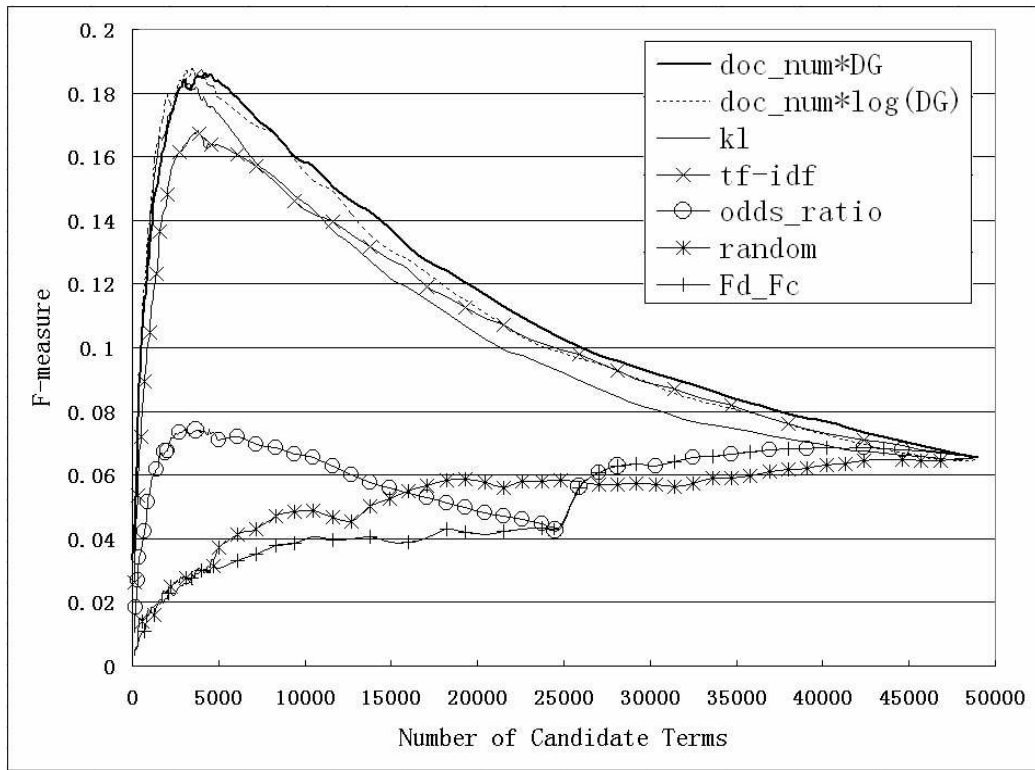


Figure 3.3. F-measure curves for different weighting methods – Astronomy / Data basis 1

In Figure 3.4, we enlarge the beginning part of the curves in Figure 3.3 by setting the range of x between 0 and 4500. The random, $F_d F_c$ and Odds Ratio curves are not displayed in this figure. It is easy to see that in the initial phase, the $\text{doc_num} \cdot \log(DG)$ curve performs the best, exceeding all other curves, while TF-IDF has the worst performance. After 4000 candidate terms, however, $\text{doc_num} \cdot DG$ takes the highest

position. KL outperforms $\text{doc_num} \times \text{DG}$ within the first 2000 candidate terms. After 3500, it will go all the way down as shown in Figure 3.3.

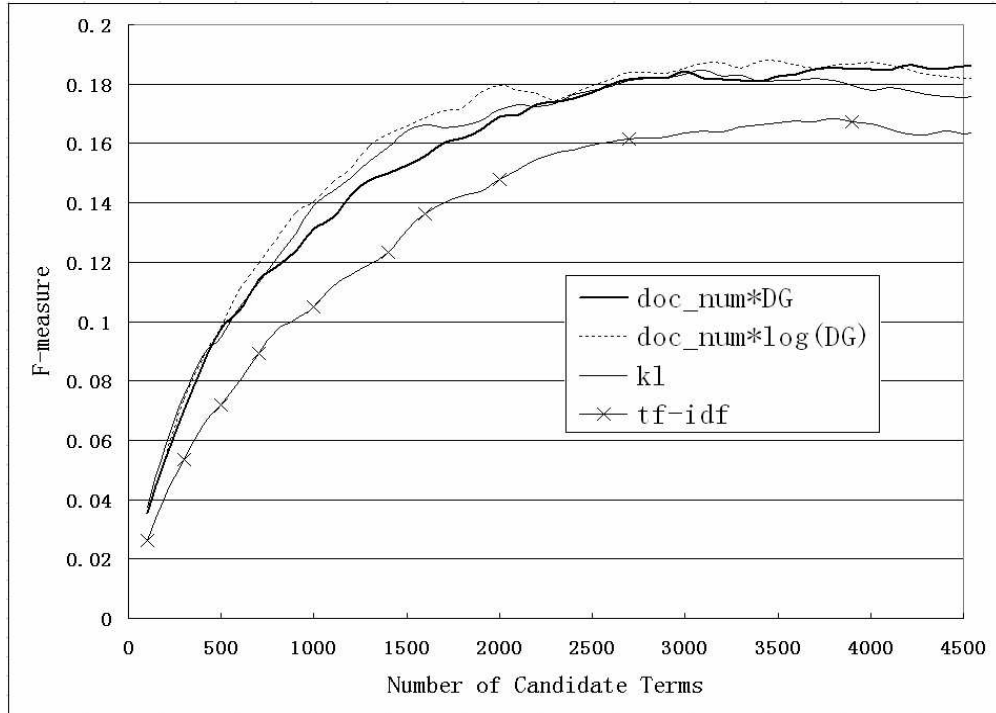


Figure 3.4. F-measure curves in a smaller range – Astronomy / Data basis 1

An interesting phenomenon from the above analysis is that all weighting approaches outperforming $\text{doc_num} \times \text{DG}$ within the initial phase (i.e. $\text{doc_num} \times \log(\text{DG})$, KL) tend to bias noise more strongly by applying the log function to the specificity factor. Since the data bases we take for the experiment are the first whole page of each URL contained in web directories, they are likely to contain much web page specific noise, such as “Click”, “Page”, “mail”, “contact” etc.. Because this noise is usually rather general, it may have a large document distribution or a high frequency in a domain. Thus, they will tend to have large topicality weight, even if their specificity is small when the log function is not used. Weighting approaches like KL and $\text{doc_num} \times \log(\text{DG})$ can better filter out such noise by

embedding the specificity factor in a log function, so that the value of topicality becomes negative if the value of specificity is smaller than 1.

In contrast, for a more specific data basis as in the second step comparison in Table 3.1, where only keywords and descriptions in a web page are used, there exists only a little noise. In this case, $\text{doc_num} \cdot \log(\text{DG})$ and KL may not work so well in the initial phase. In order to assess this assumption, we conduct a second experiment based on the candidate terms in Table 3.1, second step comparison, which is henceforth called the data basis 2. In this data basis, the number of candidate terms is reduced to 7139 and the overlap with the gold standard is only 19%

Figure 3.5 shows the result with the full range of x .

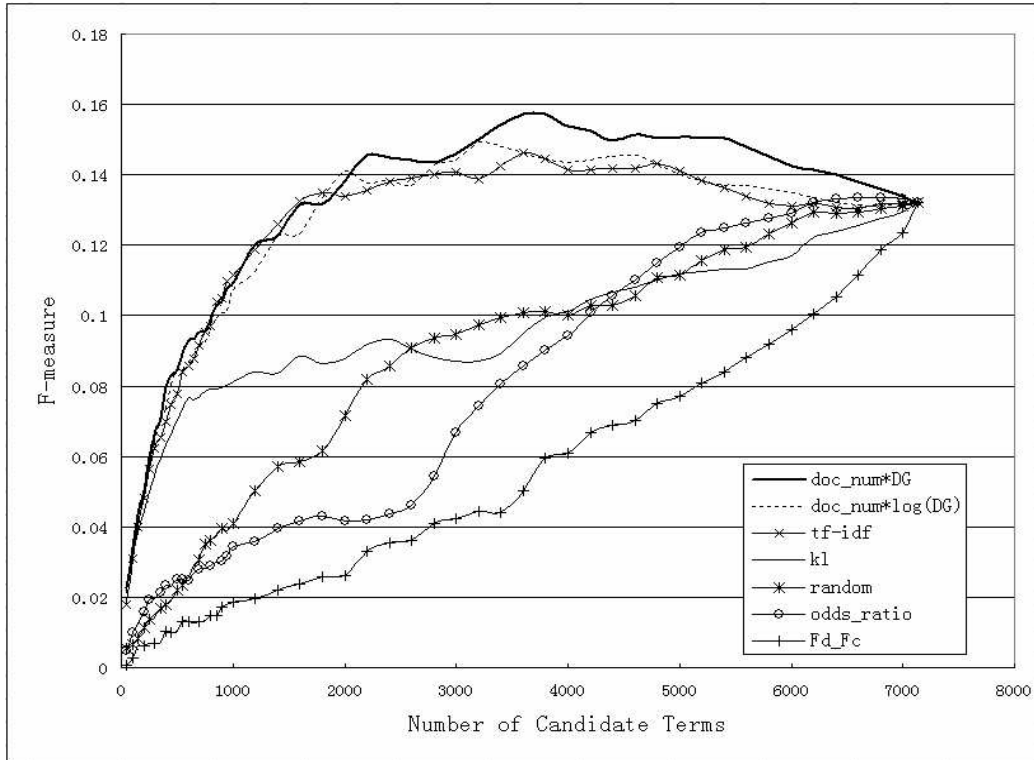


Figure 3.5. F-measure curves – Astronomy / Data basis 2

Similar to the last experiment, the two curves of our approach, especially the $\text{doc_num} \cdot \text{DG}$, lie mostly above other curves. In contrast, the KL measure performs much

worse here. It first rises as quickly as other good curves in a small range at the beginning, then it promptly loses its power and goes very flat (even dropping below the random curve in some ranges). It is interesting to see that TF-IDF works much better on this smaller data basis. It is able to keep pace with our curves, in fact, it is almost completely identical to the $\text{doc_num} \cdot \log(\text{DG})$ curve. The very good performance of the random curve is also noticeable. It lies even steadily above the Odds Ratio and Fd_Fc curves – another evidence for the low content of noise in the data basis, which generally increase the chance for the candidate terms of matching the gold standard.

Figure 3.6 shows the result on the data basis 2 for a smaller range.

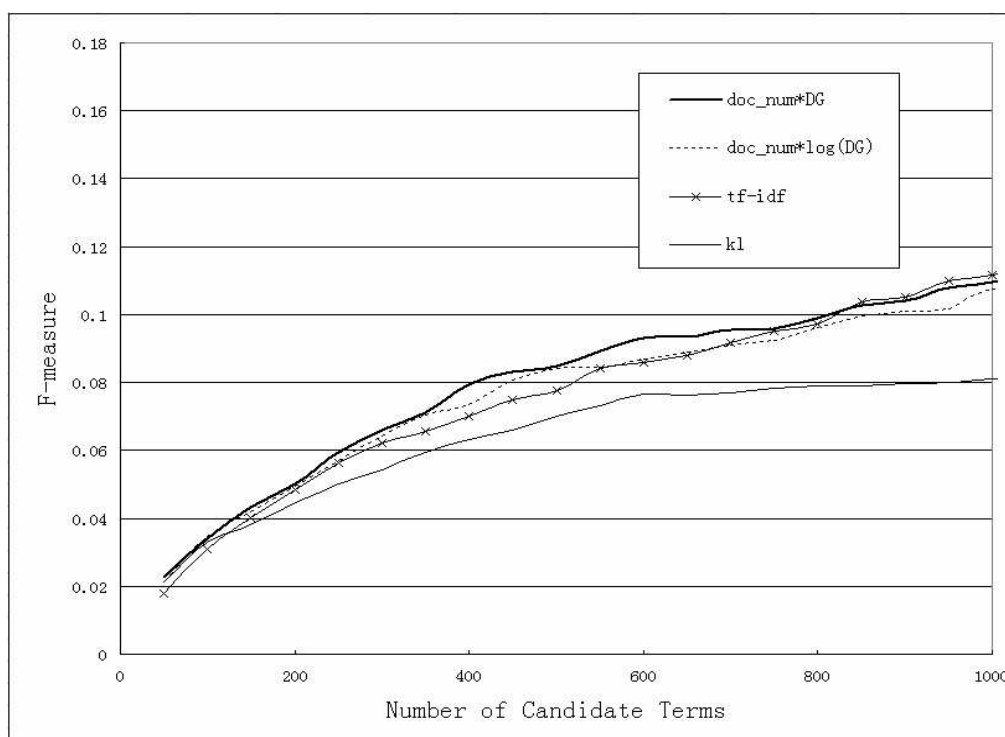


Figure 3.6. F-measure curves in a smaller range – Astronomy / Data basis 2

In addition, we have also evaluated some variations of our weighting methods, particularly some alternatives for calculating term representativeness besides doc_num , such as term frequency and a multiplication of term frequency and doc_num . Evaluation

results show that, when combined with $\log(DG)$, doc_num is the best measure for computing term representativeness.

3.3.5 Automatic Evaluation on a Second Domain

For checking the consistency of the experimental results, we have carried out another automatic evaluation for a second target domain: the domain of construction, which covers the fields of both architecture and civil engineering.

As for astronomy, a gold standard thesaurus⁸ of the domain construction is downloaded and web data are crawled from the corresponding web directories. This construction thesaurus, with a size of 15346 concepts, is much larger than the astronomy thesaurus, which has only 2900 concepts. In contrast to the astronomy thesaurus, which has strict domain-relevant terminologies, the construction thesaurus contains a number of concepts that seem very general and are not directly related to the construction domain, such as “Africa”, “Crime”, “Increase”, “Effort”, “Grammar” etc. The raw data crawled from the home page of each web site in the web directories has a size of 209 megabytes, resulting in 79501 candidate terms (after POS and $freq > 2$). These candidate terms are capable of covering nearly 1/3 concepts in the construction thesaurus (5357 of 15346), although their total number is only 1.6 times larger than in the astronomy corpus (data basis 1) and the construction thesaurus is about 5 times larger than the astronomy thesaurus.

By applying the same reference domains as for astronomy, we can compare different weighting approaches in the construction domain in Figure 3.7.

The result is somewhat similar to that in the second astronomy experiment, which is carried out on a smaller data basis (Astronomy/Data basis 2) with little noise. While the $doc_num * DG$ measure still provides the best overall performance, TF-IDF performs also

⁸ Canadian Thesaurus of Construction Science and Technology <http://irc.nrc-cnrc.gc.ca/thesaurus/toc-thesaurus.html>

very well. With a relatively large distance beneath them are the curves of $\text{doc_num} \times \log(\text{DG})$ and KL. Odds ratio and Fd_Fc are still the worst weighting approaches for extracting domain topical concepts due to the lack of a representativeness calculation. It is worth noting that the random curve in this figure, lying largely above odds ratio and Fd_Fc, performs as well as the random curve in the second astronomy experiment, which is much better than that in the first astronomy experiment.

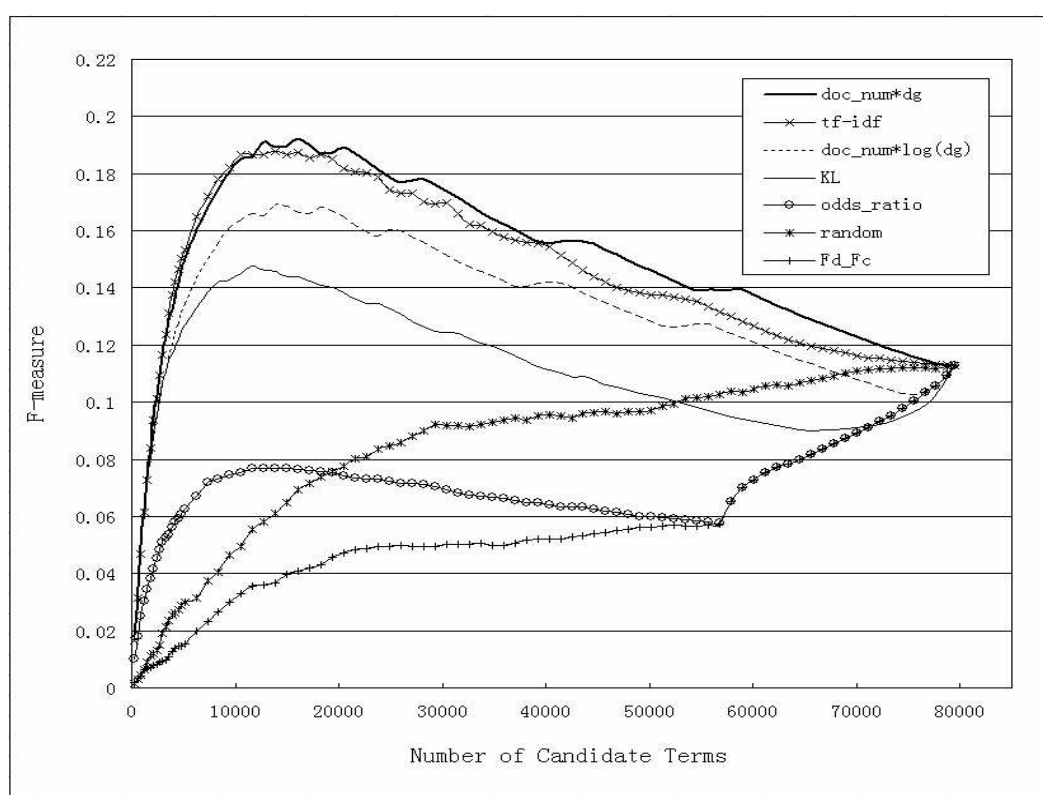


Figure 3.7. F-measure curves for different weighting methods – Construction

As indicated above, the construction thesaurus has a large size with many general concepts that are not directly related to the construction domain. This leads to a “Gold Standard Matching” of many candidate terms, which would be regarded as noise if a stricter gold standard thesaurus were applied. Due to this reason, the similarity between the result in this experiment and that in the second astronomy experiment is obvious:

while in this experiment, the web data is still “noisy” due to the usage of the whole first page of each web site, the quality of the thesaurus is reduced, which has a similar effect as reducing data noise while keeping the strictness of the thesaurus. This also explains well the good performance of the random and TF-IDF curves. Weighting measures like $\text{doc_num} \cdot \log(\text{DG})$ and KL, which bias noise more strongly by using the log function, have to suffer a performance reduction by using such a “noisy” thesaurus as the gold standard – many general terms contained in the gold standard tend to appear often in other reference domains and are thus likely to be falsely assigned a very low topicality weight by $\text{doc_num} \cdot \log(\text{DG})$ and KL, resulting in a lower value of precision and recall. However, even under this circumstance, where a distinct disadvantage for noise sensitive weighting approaches is shown, $\text{doc_num} \cdot \log(\text{DG})$ is still able to outperform the KL measure, which is consistent with the results in previous experiments.

3.4 MANUAL EVALUATION AND EXAMPLES

In the previous sections, we automatically evaluated different weighting approaches by comparing with gold standards.

However, as mentioned before, no gold standard can cover all topical terms in a domain. In the domain of astronomy, for example, the thesaurus we have used as gold standard contains only class names. It does not contain proper names like the names of individual planets (e.g. Jupiter), the names of famous organizations (e.g. NASA) or the names of famous astronomers (e.g. Kepler, Galileo). Such information may be also interesting for applications like ontology engineering and query expansion.

In order to investigate how the different weighting approaches behave upon those topical terms not contained in the gold standard, we have performed a manual evaluation on the astronomy domain and assessed the first 200 terms of each important weighting approach on the data basis 1.

Our judging criterion also accepts important proper names in astronomy as topical terms. These proper names form the biggest set of topical terms outside the gold standard. They include for example “NASA”, “JPL”, “Hubble Space Telescope”, “Galileo”, “Einstein“, “Apollo”, “Kepler”, “Leonids”, “Hale Bopp”, “ESA” and the name of planets like “Saturn”, “Jupiter”, “Pluto”, “Mars” etc.

Other terms we find interesting in the astronomy domain may also be judged as topical terms. Examples are: “observatory”, “binoculars”, “planetarium”, “spacecraft“, “Big Bang”, “Total Solar Eclipse” (the gold standard only contains the term “Solar Eclipse”) etc.

Examples of terms that are judged as non-topical are: “click”, “site”, “http”, “contact”, “news”, “welcome”, “work”, “way”, “people”, “note”, “list”, “visit” etc.

The results of the manual evaluation are shown in Table 3.4.

Table 3.4. Manual evaluation of the first 200 candidate terms in the astronomy domain

	Topical Terms in first 100 terms	Topical Terms in first 200 terms
doc_num*log(DG)	97	190
KL	91	181
doc_num*DG	89	166
TF-IDF	61	114

The results in this table confirm our conclusion of the automatic evaluation using the gold standard. In the “beginning phase”, doc_num*log(DG), KL and doc_num*DG have similar performance, where doc_num*log(DG) is with a slight difference the best method. In contrast, TF-IDF supplies too much noise in its top terms – 39 non-topical terms in top 100 terms, 13 times higher than that of doc_num*log(DG).

For a better illustration, we list the top 30 candidate terms weighted respectively by $\text{doc_num} \times \log(\text{DG})$ and TF-IDF in Table 3.5 for the domain astronomy. The “Match” column indicates if the corresponding candidate terms are matched in the gold standard, where “M” means matched and “U” unmatched. The candidate terms that we manually judged as NOT topical are signified with larger italic fonts. It is obvious that $\text{doc_num} \times \log(\text{DG})$ is able to provide much more topical concepts in the top candidate terms.

3.5 CONCLUSION

In this chapter, we have discussed how to effectively extract domain-topical concepts from web directories. We have shown that web directories are a good source for the task at hand. A new concept extraction method ($\text{doc_num} \times \text{DG}$) together with a more noise sensitive variation ($\text{doc_num} \times \log(\text{DG})$) are proposed to better weight term topicality. By employing gold standards, we evaluated our methods and other concept extraction approaches automatically in two different domains. Experimental results show that $\text{doc_num} \times \text{DG}$ is the most stable topicality weighting method, which achieves the best performance for a wide range of candidate terms and different data basis. If more noise is expected in the data basis, $\text{doc_num} \times \log(\text{DG})$ will perform better for a small range of top candidate terms.

Table 3.5. The top 30 candidate terms weighted by $\text{doc_num} \times \log(\text{DG})$ and TF-IDF in the astronomy domain. The column “Match” indicates if the corresponding terms are matched in the gold standard, with “M” for Matched and “U” for Unmatched.

Doc_num*log(DG)		TF-IDF	
Term	Match	Term	Match
astronomy	M	Earth	M
NASA	U	Sun	M
Moon	M	astronomy	M
Planets	M	Mars	U
Stars	M	Moon	M
Mars	U	Time	M
solar system	M	space	M
telescope	M	Universe	M
sky	M	NASA	U
Sun	M	Stars	M
Jupiter	U	<i>information</i>	U
Universe	M	Images	M
Earth	M	<i>click</i>	U
Saturn	U	telescope	M
comets	M	sky	M
Venus	U	<i>site</i>	U
telescopes	M	<i>Links</i>	U
orbit	M	Planets	M
galaxies	M	solar system	M
space	M	Jupiter	U
observatory	U	<i>page</i>	U
Astronomers	M	Science	U
star	M	<i>http</i>	U
galaxy	M	star	M
planet	M	<i>SEARCH</i>	U
spacecraft	U	planet	M
<i>SEARCH</i>	U	<i>University</i>	U
comet	M	light	M
Pluto	U	<i>contact</i>	U
Uranus	U	<i>Home</i>	U

CHAPTER 4 STATISTICAL RELATIONSHIP DETERMINATION

Assuming that the topical terms in a domain specific text corpus are already identified by an appropriate term weighting method as introduced in the last chapter, we discuss in this chapter how to effectively determine the relationships between these terms using statistical approaches.

Statistical approaches for relationship determination between terms are usually based on a notion of common context of terms, which is normally quantified by means of a similarity measure that compares the individual contexts of terms with their common context. Large similarity value indicates a close relationship between terms.

In this chapter, we first introduce a formalization of similarity measures in terms of conditional probabilities in Section 4.1, where the mutual conditional probability determines a “Generally Related” relationship, and individual conditional probabilities with unbalanced values determine a “Broader/Narrower” relationship. In Section 4.2 and 4.3 we formally define two main notions of context – the occurrence context and the content context, and provide a detailed analysis on their respective characteristics. In Section 4.4 we introduce a new notion of common context, and combine different notions of common context to arrive at a combined similarity measure to precisely determine the “Generally Related” relationship. In Section 4.5 we discuss how to apply the unbalanced individual conditional probabilities from the mutual conditional probability model on the premise of a close “Generally Related” relationship to determine the “Broader/Narrower” relationship. In Section 4.6 we present a series of experiments automatically evaluating the existing similarity measures and compare them with our methods. The experiments are carried out for two target domains: astronomy and construction. Several important principles are further illustrated by examples in Section 4.7. Section 4.8 draws conclusions.

4.1 CONDITIONAL PROBABILITY MODEL

In this section, we propose a formalization for measuring the similarity between terms based on conditional probabilities, which can be applied to various notions of context and commonality of context.

The simple intuition behind our approach is that the degree of relationship between two terms depends on how strongly one term implies the other, or more generally, how likely it is that the individual contexts imply a common context. This can be formalized as follows: Let t_1 and t_2 be two terms. The conditional probability $P(t_1|t_2)$ measures the probability that term t_1 occurs given that term t_2 occurs. This probability can be determined in the usual way by comparing the probability $P(t_1 \cap t_2)$ for a common context of t_1 and t_2 , denoted as $P(t_1, t_2)$, with the probability $P(t_2)$ for a context of t_2 with or without t_1 .

$$P(t_1 | t_2) = \frac{P(t_1, t_2)}{P(t_2)}$$

To measure how strongly the two terms imply each other, the conditional probabilities in both directions can be multiplied with each other to form a mutual conditional probability:

$$rel(t_1, t_2) = P(t_1 | t_2) \cdot P(t_2 | t_1) = \frac{P(t_1, t_2)^2}{P(t_1) \cdot P(t_2)}$$

Note that $P(t_1, t_2)$, $P(t_1)$ and $P(t_2)$ can be estimated in different ways according to the various notions of context and common context.

This model assigns a relatedness weight for each pair of terms to measure how likely they are generally related – we will henceforth refer to this kind of relationship as the “*Generally Related*” relationship. A large value of the mutual conditional probability can be achieved in two cases: 1. By balanced conditional probabilities, i.e. both $P(t_i|t_j)$ and $P(t_j|t_i)$ are large. 2. By unbalanced conditional probabilities, i.e. one conditional

probability is very large; another is relatively small. While the first case clearly indicates a symmetric mutual relationship between two terms, the second case suggests an asymmetric relationship, which is similar to the “subsumption” principle proposed in [76] attempting to capture the “Broader/Narrower” relationship. In section 4.5, we will discuss in detail how to use the unbalanced conditional probabilities on the premise of a close “Generally Related” relationship to determine a more reliable “Broader/Narrower” relationship.

4.2 OCCURRENCE CONTEXT

4.2.1 Formal Definitions

Occurrence context is one of the most important context types, widely applied in numerous early and recent approaches [73][25][13]. The occurrence context of a target term t can be defined as the set of text segments containing t , without taking into account the content of the text segments. A text segment may be a document or a part of a document, e.g. a paragraph, a sentence, or a window surrounding the term with a certain size. Large text segments, such as entire documents, constitute rather unspecific contexts, which do not form a reliable basis for deciding about the relationship of terms. Thus, the size of text segments is usually rather small [80][82].

For a formal definition of text segment, we first distinguish between a term – denoted as t , and an occurrence of a term – denoted as o , because a term may occur in more than one document and in a certain document, it may occur in different places. We then use $\text{term}(o)$ to denote the actual term occurring at o . Accordingly, a text corpus can be both considered as a set of terms, denoted as C_t , and a set of term occurrences, denoted as C_o . The distance between two occurrences o_1 and o_2 , denoted by $\text{dist}(o_1, o_2)$, is always

defined within a document, as the number of term occurrences occurring between o_1 and o_2 .

In our work, a text segment of a term occurrence $o \in C_o$ is defined as a window of size n surrounding o , with n being empirically set to 20, which achieved the best performance in exploratory tests.

$$\text{textseg}(o, n) := \{o' \mid o' \in C_o \wedge \text{dist}(o, o') \leq n\}$$

A text segment textseg with respect to two occurrences o_1 and $o_2 \in C_o$ is defined as:

$$\text{textseg}(o_1, o_2, n) := \{o' \mid o' \in C_o \wedge \text{dist}(o_1, o_2) \leq n \wedge \text{dist}(o_1, o') \leq n \wedge \text{dist}(o_2, o') \leq n\}$$

The occurrence context of a term $t \in C_t$ is then defined as:

$$\text{context_occ}(t, n) := \{\text{textseg}(o, n) \mid o \in C_o \wedge \text{term}(o) = t\}$$

Notice that the purpose of defining a text segment as a set of term occurrences is to enable a convenient and consistent definition of all kinds of context types, including the other two context types which will be introduced in the next sections. For defining the occurrence context in this section, however, the actual content of a text segment is ignored. A text segment will be therefore considered as a single element rather than a set of term occurrences, and the occurrence context context_occ is a set of elements rather than a set of sets.

The probability of a term t occurring in some text segments can be then estimated from the relative number of such text segments:

$$P_{\text{occ}}(t) = \frac{|\text{context_occ}(t)|}{N}$$

where N denotes the number of all text segments with a size of n in the corpus. Notice that, if the text segments of a term t can overlap, $|\text{context_occ}(t)|$ will be identical to the absolute term frequency of t .

The common context of two terms t_1 and $t_2 \in C_t$ can be defined as:

$$\text{context_occ}(t_1, t_2, n) := \{ \text{textseg}(o_1, o_2, n) \mid \begin{array}{l} t_1, t_2 \in C_t \wedge \\ o_1, o_2 \in C_o \wedge \\ \text{term}(o_1) = t_1 \wedge \\ \text{term}(o_2) = t_2 \end{array} \}$$

Again, the probability of term t_1 occurring with t_2 in a common occurrence context can be estimated from the relative number of common text segments.

$$P_occ(t_1, t_2) = \frac{|\text{context_occ}(t_1, t_2)|}{N}$$

The mutual conditional probability of two terms can then be calculated by:

$$\begin{aligned} rel_occ(t_1, t_2) &= P_occ(t_1 | t_2) \cdot P_occ(t_2 | t_1) \\ &= \frac{|\text{context_occ}(t_1, t_2)|^2}{|\text{context_occ}(t_1)| \cdot |\text{context_occ}(t_2)|} \end{aligned}$$

The square root of rel_occ corresponds to the cosine distance which will be introduced in the next section. For a better comparison with existing measures, we also take the square root for measuring the similarity of two terms by their occurrence context:

$$rel_sqrt_occ(t_1, t_2) = \frac{|\text{context_occ}(t_1, t_2)|}{\sqrt{|\text{context_occ}(t_1)| \cdot |\text{context_occ}(t_2)|}}$$

4.2.2 Similarity Measures Based on Occurrence Context

There exist numerous occurrence context based similarity measures for relationship determination, including Cosine coefficient (COS), Dice coefficient (DICE), Jaccard coefficient (JAC) [70], pointwise mutual information (PWI) [18], X^2 -test (CHI), Yule's coefficient of colligation Y (YY) [29] etc..

All these measures can be formalized as functions of four possible combinations of term pairs. They are usually described using a 2x2 contingency table as in Table 4.1 [78][88][80][17], where t_i and $\neg t_i$ represent the presence and absence of term t_i in a text segment respectively; f_{t_1,t_2} denotes the frequency of co-occurrence of t_1 and t_2 , $f_{\neg t_1,t_2}$ and $f_{t_1, \neg t_2}$ the frequency when one term occurs without the other, and $f_{\neg t_1, \neg t_2}$ the frequency when neither term occurs; $N_f = f_{t_1} + f_{\neg t_1} = f_{t_2} + f_{\neg t_2}$

Table 4.1. Contingency table for combinations of term pairs t_1 and t_2

	t_1	$\neg t_1$	
t_2	f_{t_1,t_2}	$f_{\neg t_1,t_2}$	f_{t_2}
$\neg t_2$	$f_{t_1, \neg t_2}$	$f_{\neg t_1, \neg t_2}$	$f_{\neg t_2}$
	f_{t_1}	$f_{\neg t_1}$	N_f

Table 4.2 lists some important similarity measures based on the contingency table. Instead of using the number of term occurrences, the similarity measures in Table 4.2 apply the number of contexts for similarity calculation, which simplifies term frequencies to binary form with a value of either 0 or 1, representing absence or presence of the term in the context. This simplification facilitates an intuitive comparison of different similarity measures.

It is clear that f_{t_1,t_2} , which represents co-presence of terms, makes the most important positive contribution to almost all similarity measures. In some measures like CHI and YY, absence of both terms, i.e. $f_{\neg t_1, \neg t_2}$, is also considered as positive evidence for commonality. As pointed out in [78], however, co-absence cannot be reliably applied in domains with sparse data, which is the case in the task of corpus based term relationship determination. This argument is well supported by experiments carried out in many corpus based approaches as shown in [62][80] [17].

Table 4.2. Similarity measures

COS	$\frac{f_{t_1, t_2}}{\sqrt{f_{t_1}} \times \sqrt{f_{t_2}}}$
DICE	$\frac{2 \times f_{t_1, t_2}}{f_{t_1} + f_{t_2}}$
JAC	$\frac{f_{t_1, t_2}}{f_{t_1} + f_{t_2} - f_{t_1, t_2}}$
PMI	$\log \left(\frac{N_f \times f_{t_1, t_2}}{f_{t_1} \times f_{t_2}} \right)$
CHI	$\frac{N_f \times (f_{t_1, t_2} \times f_{\neg t_1, \neg t_2} - f_{t_1, \neg t_2} \times f_{\neg t_1, t_2})}{f_{t_1} \times f_{t_2} \times f_{\neg t_1} \times f_{\neg t_2}}$
YY	$\frac{\sqrt{f_{t_1, t_2} \times f_{\neg t_1, \neg t_2}} - \sqrt{f_{t_1, \neg t_2} \times f_{\neg t_1, t_2}}}{\sqrt{f_{t_1, t_2} \times f_{\neg t_1, \neg t_2}} + \sqrt{f_{t_1, \neg t_2} \times f_{\neg t_1, t_2}}}$

Similarity measures give similarity weights to each pair of terms to determine their relationships. If the term pairs are sorted according to their similarity weights in descending order, different similarity measures may result in different ranking of these term pairs. Two similarity measures sim_1 and sim_2 will have a same effect on similarity ranking, if the following holds for all term pairs (t_1, t_2) and (t_3, t_4) :

$$\text{sim}_1(t_1, t_2) < \text{sim}_1(t_3, t_4) \Leftrightarrow \text{sim}_2(t_1, t_2) < \text{sim}_2(t_3, t_4)$$

It is easy to see that a similarity measure will have the same effect on ranking term relationships as any strictly monotonically increasing function of it. For example, the PMI measure is equivalent to the more simple measure $f_{t_1, t_2} / (f_{t_1} \times f_{t_2})$, because the log and the linear transformation $\times N$ are strictly monotonically increasing functions; and rel_occ is equivalent to rel_sqrt_occ , because the square function is also strictly monotonically increasing.

We prove in the following that DICE is a strictly monotonically increasing function of JAC, and is therefore equivalent to JAC with respect to ranking.

Note that according to Table 4.2,

$$\frac{1}{\text{DICE}} = \frac{1}{2} \cdot \left(\frac{1}{\text{JAC}} + 1 \right), \text{ hence } \text{DICE} = 2 \cdot \left(\frac{\text{JAC}}{\text{JAC} + 1} \right)$$

Differentiating DICE with respect to JAC yields

$$\begin{aligned} \text{DICE}' &= \left(2 \cdot \left(\frac{\text{JAC}}{\text{JAC} + 1} \right) \right)' \\ &= \frac{2}{(\text{JAC} + 1)^2} > 0 \end{aligned}$$

DICE is therefore a strictly monotonically increasing function of JAC.

QED

The similarity measures in Table 4.2 can also be generalized to frequencies that do not only represent the number of text segments within which a term occurs, but represent for each text segment the number of times the term occurs in the text segment. For example, with $f_{c(t_i), t_i}$ representing the frequency of t_i in a text segment $c(t_i)$ containing t_i and $f_{c(t_1, t_2), t_i}$ the frequency of t_i in a text segment $c(t_1, t_2)$ containing both t_1 and t_2 , cosine distance between t_1 and t_2 can be calculated as follows:

$$\text{COS_freq} = \frac{\sum_{c(t_1, t_2)} (f_{c(t_1, t_2), t_1} \cdot f_{c(t_1, t_2), t_2})}{\sqrt{\sum_{c(t_1)} (f_{c(t_1), t_1})^2} \cdot \sqrt{\sum_{c(t_2)} (f_{c(t_2), t_2})^2}}$$

The work of Salton [73] is one of the earliest research concerning statistical relationship determination between terms, using COS_freq as similarity measure and the whole document as text segment. The same principle is applied by Qiu and Frei [67] for constructing thesaurus for query expansion.

Jing and Croft [46] simply multiplied the frequency of a concept (noun phrase) with the frequency of a term in the text segments the concept and the term co-occur to determine their relationship. A concept was then represented by a term vector, which can be used for query expansion. The similarity between a concept and a query was computed by COS_freq. In this work, paragraphs in a document were used as text segments. If a natural paragraph is too long, it will be further divided into 3-10 sentences.

Curran stated in [26] that Dice and Jaccard have the best performance for determining relationships between terms, where Dice is easier to compute and is thus the preferred measure. Mandala [60] also applied Dice as the association measure to build a thesaurus.

Pointwise mutual information (PMI) is another widely applied similarity measure, firstly used for relationship determination between terms by Church et. al. [18]. If two terms, x and y , have probabilities of occurrence $P(x)$ and $P(y)$, then their pointwise mutual information, $I(x,y)$, is defined to be

$$I(x, y) = \log \left(\frac{P(x, y)}{P(x) \cdot P(y)} \right)$$

This measure compares the probability that x and y co-occur (the joint probability) with the probabilities of x and y occurring independently. If x and y are related to each other, then the joint probability $P(x,y)$ will be much larger than $P(x) \times P(y)$, resulting in a $I(x,y)$ much greater than 0. If there is no interesting relationship between x and y , $P(x,y)$ will approximately equal to $P(x) \times P(y)$, and thus, $I(x,y)$ approximately equals to 0. If x and y are in complementary distribution, then $P(x,y)$ will be much smaller than $P(x) \times P(y)$, resulting in a $I(x,y)$ that is much smaller than 0. When used for determining relationship between two terms in a corpus, the probabilities above can be estimated as the PMI in Table 4.2. A window with a fixed size of 5 was used as text segment.

Turney [82] tried to use PMI to find synonyms. As shown in his experiments, PMI outperformed the Latent Semantic Analysis (LSA) for synonym detection if text segments are restricted to smaller windows. The works of Terra et. al. [80] and Baroni [4] confirmed the good performance of PMI in finding synonyms. It was shown to outperform some occurrence based approaches like the CHI-test and most of the content based approaches, such as the content based cosine measure and other distributional similarity measures. We will introduce the content based approaches in more detail in the next section. However, as French et. al. argued in their paper [34], what Turney and other actually found were highly related terms, not synonyms. Real synonyms cannot be reliably detected by purely statistical approaches. Another problem with the approaches applying PMI for term relationship determination is that they did not compare PMI with other important occurrence based similarity measures like Cosine, Dice and Jaccard. Further evaluations are therefore desirable.

Among the similarity measures introduced above, COS distance is the most similar one to the `rel_sqrt_occ` measure derived from the mutual conditional probability model (They are actually equivalent to each other when using the number of contexts instead of using the number of occurrences for similarity calculation). Other similarity measures like PMI, DICE and JAC use rather different ways for using and normalizing common context for relationship determination and thus are expected to perform differently from `rel_sqrt_occ` and COS.

4.2.3 Problems of Occurrence Based Approaches

Despite of their simplicity and high efficiency, approaches based on occurrence context suffer from several well-known problems.

One problem lies in the requirement of co-occurrence. That is, two terms are only considered similar if they occur in same contexts in a text corpus for a certain number of times. The problem matters in two cases. First, due to sparsity, a text corpus (even a very

large one) can only cover a (small) part of a total vocabulary in a domain and tends to contain even less co-occurring information between the vocabularies [28]. Term pairs with interesting relationships that do not (often) co-occur in text corpus will have no chance to be highly weighted. This data sparseness problem of text corpora can be partially solved by carefully choosing a proper text corpus with reasonably wide domain coverage, such as an appropriate category in web directories (see Section 3.1 for detail information). The second problematic case is the detection of spelling variations and synonyms, such as “color” and “colour”, “astronaut” and “cosmonaut”, which are not likely to co-occur in the same context at all although they are very close terms. Such relationships can hardly ever be detected, no matter how large the underlying text corpora are. Both of these cases of co-occurrence lead to a reduction of recall in relationship determination.

Another drawback of occurrence based approaches is that they do not take the content of context into consideration. This leads to the problem that multiple co-occurrences of terms t_1 and t_2 with similar content may increase the similarity of t_1 and t_2 too strongly. In other words, if a text corpus contains many texts with similar content, occurrence based approaches may overweight term relationships in these texts, resulting in a reduction of precision. This is especially the case for news corpora, where an unusual event such as “A plane landed on a highway yesterday” tends to be reported in many news articles with similar content. Occurrence based approaches will therefore observe frequent co-occurrence of “plane” and “highway” in many contexts and give them incorrectly large similarity weight.

4.3 CONTENT CONTEXT

4.3.1 Formal Definitions

One way to overcome the problems of using occurrence context is to look at the actual terms in the context of two target terms t_1 and t_2 . The relationship between t_1 and t_2 will be calculated transitively through their common context terms. Figure 4.1 illustrates the essential principle of content based approaches.

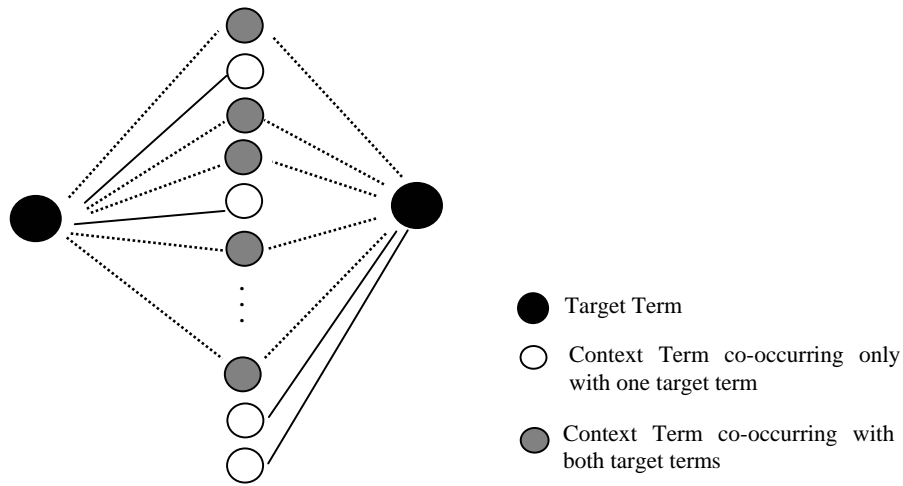


Figure 4.1. Relationship determination through the common context terms

Based on this principle, we define the content context for a term $t \in C_t$ as:

$$\text{context_con}(t, n) := \{t_{\text{con}} \mid t_{\text{con}} \in C_t \wedge \exists o, o_{\text{con}} \in C_o \text{ s. t.} \\ \text{term}(o) = t \wedge \text{term}(o_{\text{con}}) = t_{\text{con}} \wedge o_{\text{con}} \in \text{textseg}(o, n) \}$$

When using content context for relationship determination, the probability that term t occurs can be estimated from the relative number of the unique terms that occur in the content context of t .

$$P_con(t) = \frac{|context_con(t)|}{N'}$$

where N' denotes the sum of cardinalities of the content contexts of all target terms in the corpus.

The common content context of two terms t_1 and t_2 can be defined as the intersection of the individual contexts of t_1 and t_2 .

$$context_con(t_1, t_2) := context_con(t_1) \cap context_con(t_2)$$

The probability that term t_1 is related to t_2 can be estimated from the relative number of the unique terms that occur both in the content context of t_1 and in the content context of t_2 :

$$P_con(t_1, t_2) = \frac{|context_con(t_1, t_2)|}{N'}$$

The cardinalities of the individual and common content contexts can again be used to determine a mutual conditional probability:

$$\begin{aligned} rel_con(t_1, t_2) &= P_con(t_1 | t_2) \cdot P_con(t_2 | t_1) \\ &= \frac{|context_con(t_1, t_2)|^2}{|context_con(t_1)| \cdot |context_con(t_2)|} \end{aligned}$$

By taking into account the content of a context one may also discover relationships between terms that do not co-occur frequently but each co-occurs with the same terms frequently. For example, a relationship between the terms “Earth” and “Saturn” may be inferred based on sentences such as “The Earth orbits the Sun” and “The Saturn orbits the Sun”, even if “Earth” and “Saturn” do not occur together.

4.3.2 Similarity Measures Based on Content Context

Some similarity measures in Table 4.2 such as COS, DICE, JAC and PMI can be directly applied for content context if we redefine $f_{i1, i2}$ as the number of unique common

context terms of t_1 and t_2 ; f_{t1} , f_{t2} the numbers of unique context terms of t_1 and t_2 respectively. They can also be generalized for considering real term frequencies as the occurrence based similarity measures. The COS_freq formula introduced in the last section, for example, can be directly used for content context if we let $f_{c(t_i),t_i}$ represent the frequency of a context term $c(t_i)$ in the context of t_i and $f_{c(t_1,t_2),t_i}$ the frequency of a context term $c(t_1,t_2)$ in the context of t_i , where the term $c(t_1,t_2)$ should occur with t_1 in same text segments, and also with t_2 in same (other) text segments.

In addition, there exists a series of the so-called distributional similarity measures [28], which first explore the distributions of target terms in their respective content context, and then apply probabilistic measures to compare the distributions. These measures include L1 Norm (L1), Contextual Jensen-Shannon Divergence (CJSD), Contextual Average Mutual Information (CAMI) etc., which are listed in Table 4.3, assuming t_1 and t_2 are two target terms, whose relationship should be determined and t' is a context term shared by t_1 and t_2 .

Table 4.3. Distributional similarity measures

L_1	$\sum_{t'} P(t' t_1) - P(t' t_2) $
CAMI	$\sum_{t'} P(t' t_1) \cdot \log\left(\frac{P(t' t_1)}{P(t' t_2)}\right)$
CJSD	$KL(p \parallel q) = \sum p \cdot \log\left(\frac{p}{q}\right)$ $AVGP = \frac{P(t' t_1) + P(t' t_2)}{2}$ $CJSD = KL(P(t' t_1) \parallel AVGP) + KL(P(t' t_2) \parallel AVGP)$

Due to their ability in better solving the data sparseness problem, content based approaches have been shown to have rather good performance in some applications like

language modelling [28] and Information Retrieval [36][86]. However, there is no evidence that they also work well for relationship determination in automatic construction of concept structures. In contrast, evaluations for the task of automatic synonym detection (actually highly related terms) [80] showed that content based approaches generally perform worse than the occurrence based ones. Baroni et. al. [4] also showed that occurrence based PMI performs much better than content based COS in finding related terms.

In our experiments, purely content based approaches do not perform well either. In fact, many term relationships highly weighted by these approaches are not interesting at all, resulting in a relatively low precision compared to occurrence based approaches

4.3.3 Problems of Content Based Approaches

We identify three possible reasons which may explain the poor performance of content based approaches in our task: word sense ambiguity, untopical context terms and random overlapping of context terms. While the first one should not matter much as we restrict our work on domain specific thesauri, the second one should be resolvable with a little more effort, and we see no possibility of solving the third one with purely content based approaches. In the following we will explain these three points in more detail.

– Word Sense Ambiguity

Word sense ambiguity, referring to the fact that a word may have multiple meanings in different domains, is a severe problem in constructing general purposed thesauri. In a generic text corpus, contexts corresponding to different meanings of a word can hardly be distinguished, which may lead to an incorrect relationship determination. In our work on automatic construction of domain specific concept structures, however, we restrict relationship determination to domain specific text collections, where words tend to have more restricted meaning than in general language. For example, in the astronomy domain, the term “eclipse” refers very probably only to a special astronomical phenomenon,

rather than to the open source development toolkit for Java. The contexts of a word will therefore be almost always related to a single meaning of a word.

– Untopical Context Terms

In content based approaches, terms occurring in the contexts of target terms play a key role for relationship determination. However, for construction of domain specific thesauri, untopical context terms usually do not carry any significant information of target terms. Many of these context terms will also occur with both target terms. But because they do not carry significant information, they cannot be used for inferring a meaningful relationship. For example, in the sentences: “This is a beautiful photograph” and “This is a beautiful planet”, two target terms “planet” and “photograph” have four common context terms “This”, “is”, “a” and “beautiful”, which are not topical in the domain of astronomy – the first three words are actually stop words. These context words could lead to a relationship between “planet” and “photograph”, even if this relationship is rather remote.

This problem can be solved by weighting terms surrounding target terms according to their topicality in a target domain (See Chapter 3 for more information). Only topical terms will be chosen to constitute content based context.

– Random Overlapping of Context Terms

Although restricted in a certain domain, the meaning of a target term may still have multiple aspects, i.e. different domain specific topics in which the target term may be involved. The more generic a target term is, the more aspects it tends to have. Two target terms having close relationship must have one or more non-trivial common aspects.

Conventional content based approaches use the number of total overlapped context terms as indication of existing common aspects and compare it to the numbers of context terms of the respective target terms to see if the common aspects are non-trivial (c.f. the formula of mutual conditional probability in Section 4.3.1).

However, as we observe, context terms of both target terms do not only overlap in the common aspects, they may also randomly overlap in non-common aspects. Figure 4.2 depicts an example of two target terms “sun” and “photograph” in the domain astronomy. We assume both terms have non-ambiguous meaning in the domain, and only the topical context terms of both target terms are considered.

As both of the target terms are rather generic, they can be involved in many different topics, and thus have many aspects of meaning. We display a part of these aspects in the Figure 4.2 with white squares. The shaded areas in the white squares represent the overlapping context terms of two target terms. As marked by the oval dotted line, the two target terms “sun” and “photograph” share only one small common aspect, namely “photograph of sun”. However, the common context terms of the two target terms do not only occur in the common aspect. They are also scattered in many other non-common aspects. For example, the context terms “Mars”, “Jupiter”, “Saturn”, “comets”, “satellites”, “polar light” etc., which co-occur with the target term “sun” in the aspects “solar system” or “impact on earth”, may also co-occur with another target term “photograph” in other aspects like “photograph of planets”, “photograph of comets” or “photograph of astronomical impact on earth”.

Since these randomly overlapping context terms in the non-common aspects cannot be effectively distinguished from the meaningfully overlapping context terms in the common aspects by using purely content based approaches, one has to rely on the total number of overlapping context terms, and use it to indicate the existence of common aspects, which, however, tends to overestimate the relationships between target terms that have many randomly overlapping context terms.

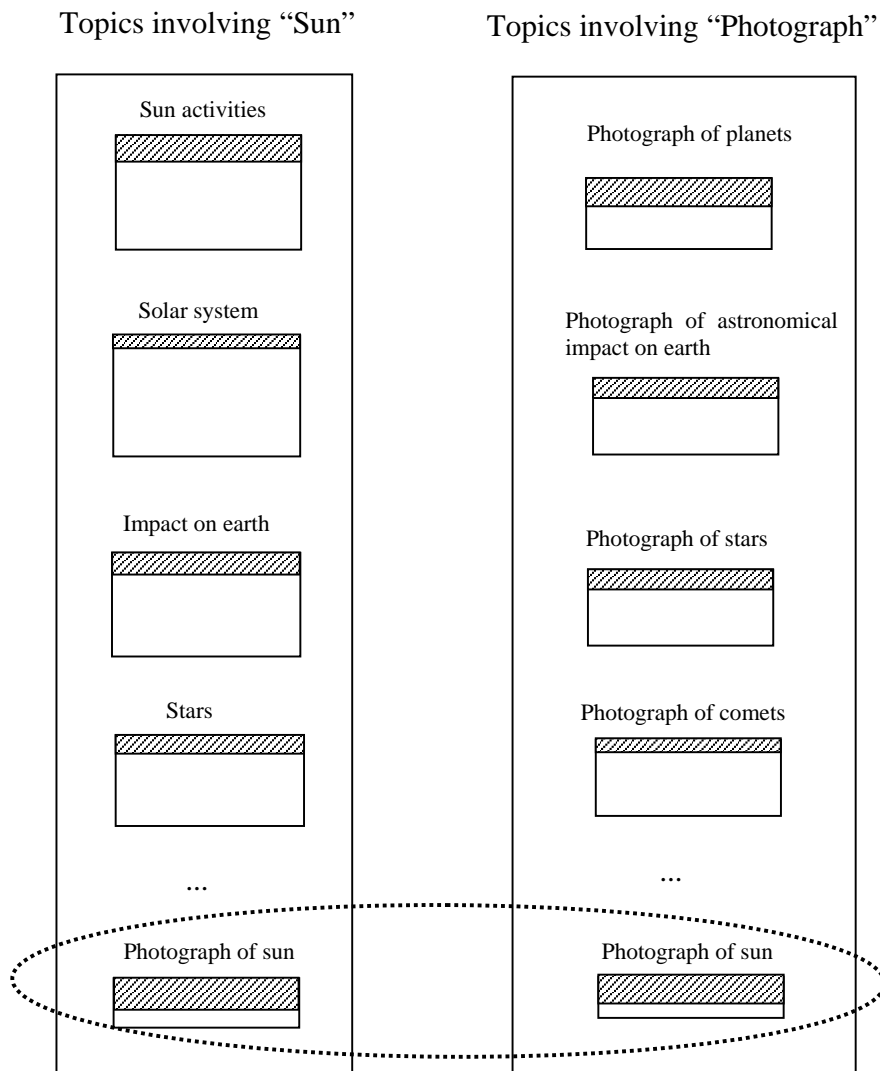


Figure 4.2. Topics involving "Sun" and "Photograph"

4.4 COMBINING OCCURRENCE CONTEXT WITH CONTENT CONTEXT

To overcome the problems of occurrence context and content context we combine them to form a notion of common context based on co-occurrence and common content:

$$\begin{aligned} \text{context_occ_con}(t_1, t_2, n) := \{ t_{\text{con}} \mid & t_{\text{con}} \in C_t \wedge \\ & \exists o_1, o_2, o_{\text{con}} \in C_o \text{ s. t.} \\ & \text{term}(o_1) = t_1 \wedge \\ & \text{term}(o_2) = t_2 \wedge \\ & \text{term}(o_{\text{con}}) = t_{\text{con}} \wedge \\ & o_{\text{con}} \in \text{textseg}(o_1, o_2, n) \} \end{aligned}$$

From one point of view, this is a content context that considers the content in the common text segments, while from another point of view, this is an occurrence context, which requires that two target terms, i.e. t_1 and t_2 , co-occur in same text segments.

4.4.1 Comparison with Purely Content Based Common Context

Compared with purely content based common context, the combined common context considers only context terms t_{con} , which co-occur with t_1 and t_2 , and are thus more likely to stand for common aspects of these terms.

Figure 4.3 illustrates the content based (common) contexts and the combined common context of t_1 and t_2 , where the combined common context is always a subset of the content based common context.

Carrying the sun-photograph example in the last section further, the terms in the context of co-occurrences of “sun” and “photograph” are more likely to be concerned with both target terms, such as terms about special devices or techniques for photographing of the sun. Other terms such as “Saturn”, “Mars” and “polar light”, which appear in different aspects of different target terms but randomly overlap, are not likely to be contained in the combined common context.

After successfully distinguishing common aspects by using the combined common context, we still need to check if the common aspects are non-trivial to the respective target terms.

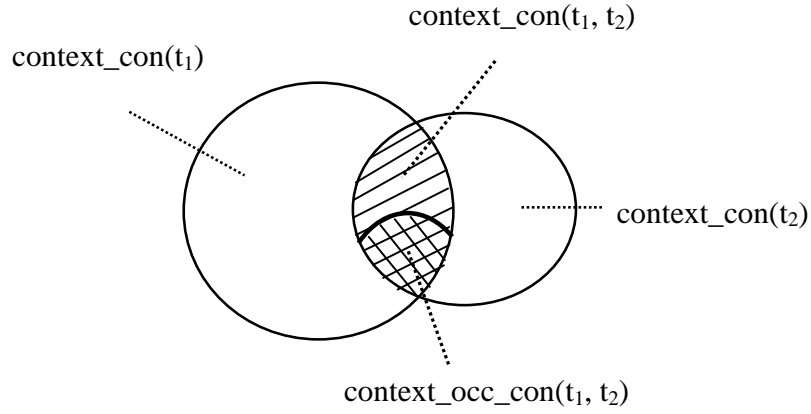


Figure 4.3. Different types of context of common context

One way to achieve this task is to normalize the cardinality of the common context by the cardinalities of the individual content context of target terms. This can be well calculated by the mutual conditional probability model, where the probability that a target term t occurs, i.e. $P_con(t)$, is estimated in the same way as in Section 4.3.1, and the probability that a target term t_1 is related to a target term t_2 , i.e. $P_occ_con(t_1, t_2)$, is estimated by:

$$P_occ_con(t_1, t_2) = \frac{|context_occ_con(t_1, t_2)|}{N'}$$

with N' denoting the sum of cardinalities of the content contexts of all target terms in the corpus.

The conditional probability $P_occ_con(t_1 | t_2)$ will then be calculated as

$$P_occ_con(t_1 | t_2) = \frac{P_occ_con(t_1, t_2)}{P_con(t_2)} = \frac{|context_occ_con(t_1, t_2)|}{|context_con(t_2)|}$$

$P_{occ_con}(t_2 | t_1)$ can be calculated in a similar way.

The cardinalities of the combined common context and individual content contexts can then be used to determine a mutual conditional probability:

$$\begin{aligned} rel_occ_con(t_1, t_2) &= P_{occ_con}(t_1 | t_2) \cdot P_{occ_con}(t_2 | t_1) \\ &= \frac{|context_occ_con(t_1, t_2)|^2}{|context_con(t_1)| \cdot |context_con(t_2)|} \end{aligned}$$

Again we take the square root of this probability for a better comparison with the cosine distance:

$$rel_sqrt_occ_con(t_1, t_2) = \frac{|context_occ_con(t_1, t_2)|}{\sqrt{|context_con(t_1)| \cdot |context_con(t_2)|}}$$

Another way of using the combined common context for relationship determination is to calculate the ratio between two different types of mutual conditional probabilities.

$$\frac{P_{occ_con}(t_1 | t_2) \cdot P_{occ_con}(t_2 | t_1)}{P_{con}(t_1 | t_2) \cdot P_{con}(t_2 | t_1)} = \frac{|context_occ_con(t_1, t_2)|^2}{|context_con(t_1, t_2)|^2}$$

As usual, we take the square root of the ratio for further relationship determination, which we henceforth refer to as the aspect ratio.

$$aspect_ratio = \frac{|context_occ_con(t_1, t_2)|}{|context_con(t_1, t_2)|}$$

This measure actually normalizes the cardinality of the combined common context by the cardinality of the common content context. It explicitly deals with the problem of spurious common context terms which occur in $context_con(t_1, t_2)$ but do not occur in $context_occ_con(t_1, t_2)$. It is capable of ruling out spurious relationships between generic terms that tend to have many common context terms, such as “sun” and “photograph”, whereby $context_con$ becomes large and $context_occ_con$ is relatively small. An

advantage of aspect ratio over rel_sqrt_occ_con is that the aspect ratio does not rule out meaningful relationships between a generic term and a specific term, such as “telescope” and “Ritchey Chretien Telescope”. The reason can be best illustrated in Figure 4.4, where t_1 represents a generic term and t_2 a specific one. Because a specific term generally shares less common context terms with other terms, the intersection of $\text{context_con}(t_1)$ and $\text{context_con}(t_2)$, i.e. $\text{context_con}(t_1, t_2)$, is usually small. In this case, although the combined common context $\text{context_occ_con}(t_1, t_2)$ between t_1 and t_2 is quite small, which usually leads to a low value of rel_sqrt_occ_con due to the large value of $\text{context_con}(t_1)$, the ratio between $\text{context_occ_con}(t_1, t_2)$ and $\text{context_con}(t_1, t_2)$, i.e. the aspect ratio, may still be large, which calculates the relationship between a generic term and relatively specific term more accurately.

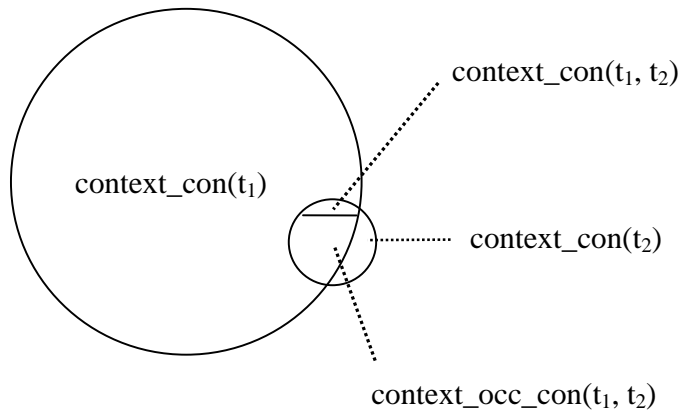


Figure 4.4. The relationship between a general term and a relatively specific term

4.4.2 Comparison with Purely Occurrence Based Common Context

Compared with the purely occurrence based common context, the combined common context considers the content of the contexts of the co-occurring target terms. The cardinality of this common context gives equal importance to all distinct context terms t_{con} used with t_1 and t_2 , no matter how often t_1 and t_2 co-occur. Thus, multiple co-

occurrences of target terms with similar content influence the similarity of these terms less strongly.

Carrying the example “A plane landed on a highway yesterday” in Section 4.3.3 further, we assume that this relatively unusual event is widely reported in many news articles. The number of such articles, which tend to have very similar content, is denoted by k , which may have a rather large value. The frequency with which the two target terms “plane” and “highway” co-occur will then also be k . No matter how large the value of k will be, since all these co-occurrences have similar context content, the number of their unique context terms will remain constant. In particular, if we define a text segment as a sentence, the set of non-stopword context terms contains only two elements “landed” and “yesterday”. The cardinality of this set will then always be two, which is more appropriate for indicating the relationship between “plane” and “highway” by eliminating redundant information carried by the repeated similar content.

To check if the common aspects – determined by the combined common context – are non-trivial to the respective target terms, we can again normalize the cardinality of the combined common context, this time by the cardinalities of the individual occurrence contexts of target terms, because we consider the combined common context here from the view of occurrence context. The relationship measure between t_1 and t_2 , i.e. $rel_sqrt_occ_con$, can thus also be calculated as:

$$rel_sqrt_occ_con(t_1, t_2) = \frac{|context_occ_con(t_1, t_2)|}{\sqrt{|context_occ(t_1)| \cdot |context_occ(t_2)|}}$$

Since this measure has the same numerator as the $rel_sqrt_occ_con$ measure presented in the last section, and the denominators of both measures, i.e. $|context_occ(t_i)|$ and $|context_con(t_i)|$, are somewhat co-related, the two measures are expected to perform similarly. However, as we notice, using $|context_occ(t_i)|$ as denominator helps better solving the “plane / high way” problem described above, where multiple co-occurrences

have similar content in their contexts. In such a case, not only the cardinality of the combined common context $|\text{context_occ_con}|$, but also the cardinalities of the individual content contexts $|\text{context_con}(t_i)|$ tend to be small because of the repeated content. This may result in a relatively large value of rel_sqrt_occ_con , which incorrectly indicates a close relationship between two target terms like “plane” and “high way”. In contrast, since multiple co-occurrences of two target terms usually cause a relatively large value of $|\text{context_occ}(t_i)|$, using $|\text{context_occ}(t_i)|$ as denominator guarantees a small value of rel_sqrt_occ_con , which correctly indicates a remote relationship between “plane” and “high way”. We therefore prefer using $|\text{context_occ}(t_i)|$ as denominator to calculate rel_sqrt_occ_con in our work.

4.4.3 A Combined Similarity Measure

In previous sections, we have discussed several similarity measures based on different types of (common) context in a framework of mutual conditional probability, including rel_sqrt_occ , rel_sqrt_con , rel_sqrt_occ_con , and aspect_ratio .

Compared to rel_sqrt_con , which suffers from low precision due to random overlapping of context terms, rel_sqrt_occ is more accurate and much simpler to implement. rel_sqrt_occ_con can reduce the too strong influence of multiple co-occurrences of target terms with similar content of context on relationship determination. It can be applied as a good complement to rel_sqrt_occ , which does not consider the content of context at all. Therefore, combining rel_sqrt_occ and rel_sqrt_occ_con will help to raise the precision of relationship determination, especially when the underlying text corpus contains many text segments with similar content, as a news corpus usually does.

As a generally well performing similarity measure, the aspect ratio tends to be especially capable of including more relationships between general terms and relatively specific terms than other similarity measures, by using the less precise common content

context context_con to normalize the more precise combined common context context_occ_con.

Now we have three individually well-performing relationship determination measures, i.e. aspect_ratio, rel_sqrt_occ_con and rel_sqrt_occ, with each of them considering different statistical evidence to compute semantic relationships. Intuitively, the more evidence about a relationship is considered, the more reliably the relationship can be calculated, which suggests that a conjunctive combination of different kinds of evidence helps to achieve the most reliable results. We therefore multiply the three measures to form a hybrid measure to achieve an optimal performance in relationship determination, as shown in the following formula:

$$rel_combined(t_1, t_2) = aspect_ratio(t_1, t_2) \cdot rel_sqrt_occ_con(t_1, t_2) \cdot rel_sqrt_occ(t_1, t_2)$$

4.5 DETERMINATION OF THE “BROADER/NARROWER” RELATIONSHIP

While the previous sections deal with the determination of symmetric relationships, we discuss in this section how to find asymmetric relationships between terms, which play an important role in building hierarchical concept structures.

As described in the chapter of related work in Section 2.1.3, Glover et. al. [38] was able to distinguish terms as “parent”, “self” and “child” by comparing their frequencies in the category with that in the whole collection. In this way, a hierarchical term structure with three layers, i.e. “parent”, “self”, “child”, could be built.

Sanderson et. al. [76] used the following conditional probabilities to determine an asymmetric “subsumption” relationship, which is expected to approximate the “Broader/Narrower” relationship in a manually built thesaurus.

$$\begin{cases} P(t_1 | t_2) \geq 0.8 \\ P(t_2 | t_1) \leq P(t_1 | t_2) \end{cases}$$

where the bracket symbol above represents an “AND” connection between two or more formulae. Term t_1 is regarded to “subsume” term t_2 , if t_2 often occurs with t_1 in same text segments (should be at least 80% of the time), and t_1 does not often co-occur with t_2 . t_1 can then be regarded as a general term broadly defining a topic, and t_2 explains a subtopic of t_1 . In Sanderson’s work, a text segment is implemented as a document. A similar principle is adopted by Lawrie [50] for hierarchical summarization of document collections. However, she implemented text segments as smaller text windows instead of using whole documents. A text window consists of k words to the left and to the right of a target term, where k should be specified by users at the time a hierarchy is created.

It is worth noting that the way of applying conditional probabilities in Sanderson’s work actually corresponds to a special case of unbalanced individual conditional probabilities in our mutual conditional probability model. As briefly discussed in Section 4.1, a relatively large $P(t_i|t_j)$ and a relatively small $P(t_j|t_i)$ may reveal an asymmetric relationships between t_i and t_j . Sanderson’s work empirically sets the threshold of the larger conditional probability to 0.8, and restricts another conditional probability to be smaller than the first one, so that the values of the two individual conditional probabilities are kept unbalanced.

For a better explanation of using unbalanced conditional probability to find a “Broader/Narrower” relationship, let us consider an example based on occurrence context (other context types follow a similar principle). Suppose t_1 is a “bigger” term that occurs very often in a domain, and t_2 a “smaller” term occurring relatively rarely in the domain. A relatively large value of $P(t_1|t_2)$ indicates that when t_2 occurs, it often occurs with t_1 together in same text segments, while a relatively small value of $P(t_2|t_1)$ indicates that t_1 also occurs with many other terms in text segments where t_2 does not occur. Such unbalanced conditional probabilities seem to well imply the existence of a

“Broader/Narrower” relationship between t_1 and t_2 , with t_1 being the general term and t_2 the specific term, because a specific term often co-occurs with its general term, but the general term tends to also co-occur with other of its specific terms. This principle is further illustrated in Figure 4.5 with the term “Linux” as t_1 and the term “SUSE” as t_2 , a well-known distribution of Linux developed in Germany. As we can see, when “SUSE” occurs, it tends to always occur with “Linux” together in same text segments. However, since there exist also other Linux distributions such as “Red Hat” or “Mandrake”, the term “Linux” can also occur with these terms in text segments without “SUSE”.

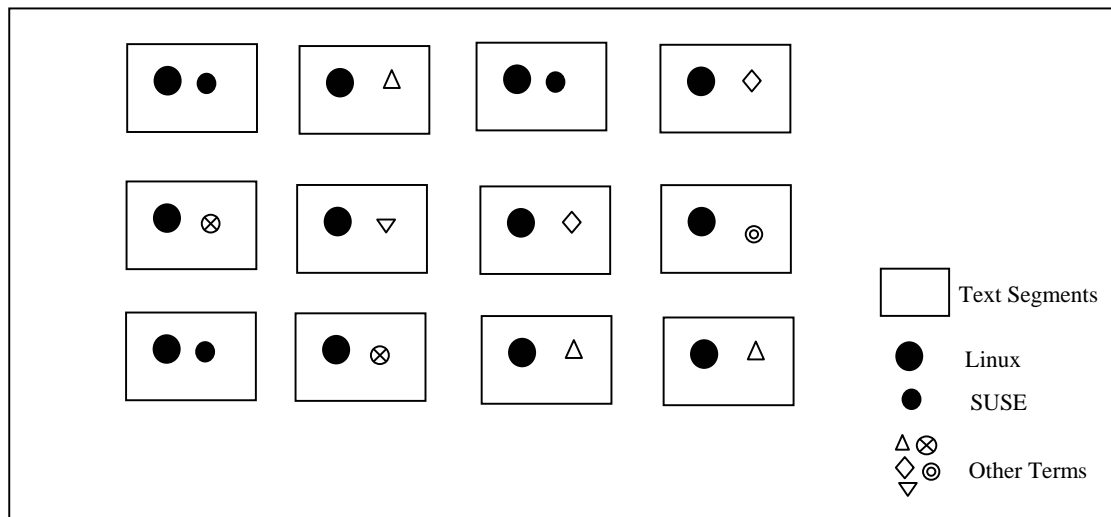


Figure 4.5. Distributions of “Linux” and “SUSE” in different text segments

As we have found out, however, considering only unbalanced individual conditional probabilities without taking into account other conditions may not always reliably determine a “Broader/Narrower” relationship. It is true that term pairs with a real “Broader/Narrower” relationship are likely to have unbalanced individual conditional probabilities. But such unbalanced individual conditional probabilities are also often observed with many term pairs with rather remote relationships. This is a severe problem because the number of such term pairs in a text corpus usually greatly exceeds the number of term pairs that have a real “Broader/Narrower” relationship. Using unbalanced

conditional probabilities alone as a criterion may therefore result in an unexpected low precision.

To address this problem, we consider the unbalanced individual conditional probabilities based on the premise of a close “Generally Related” relationship. As shown in following formulae, we first check whether the mutual conditional probability between two terms t_1 and t_2 is greater than a predefined threshold th_1 . The terms are regarded to have a close “Generally Related” relationship if the condition is satisfied. Two additional thresholds th_2 and th_3 (th_3 is usually smaller than or equal to th_2) are then used to determine the unbalanced conditional probabilities of these two terms. Note that all of these thresholds should be set empirically in experiments.

$$\begin{cases} rel(t_1, t_2) = P(t_2 | t_1) \times P(t_1 | t_2) \geq th_1 \\ P(t_1 | t_2) \geq th_2 \\ P(t_2 | t_1) \leq th_3 \end{cases}$$

Like the mutual conditional probability, the conditional probabilities $P(t_1|t_2)$ and $P(t_2|t_1)$ can also be estimated by different notions of context and common context. In Sanderson’s work [76], for example, the conditional probabilities are estimated based on occurrence context with text segments implemented as documents. In our notation, they can be described as:

$$P_{occ_doc}(t_1 | t_2) = \frac{|context_occ_doc(t_1, t_2)|}{|context_occ_doc(t_2)|}$$

$$P_{occ_doc}(t_2 | t_1) = \frac{|context_occ_doc(t_1, t_2)|}{|context_occ_doc(t_1)|}$$

where `context_occ_doc` denotes a special type of `context_occ` by using document as text segment.

Since the mutual conditional probability $rel(t_1, t_2)$ only determines a premise for a reliable application of unbalanced individual conditional probabilities, it can be estimated

independently from $P(t_1|t_2)$ and $P(t_2|t_1)$ by a different type of context and common context. For example, while the individual conditional probabilities may be estimated by document based occurrence context, i.e. `context_occ_doc`, the mutual conditional probability can be estimated by occurrence context based on smaller text segments, i.e. `rel_occ(t1,t2)`.

By restricting the determination of unbalanced conditional probabilities to closely related term pairs, term pairs that have unbalanced conditional probabilities but a remote relationship are automatically filtered out, which may help to improve the precision in determining the “Broader/Narrower” relationship to a great extent.

4.6 AUTOMATIC EVALUATION

In this section we intent to automatically evaluate the performance of some most important statistical measures with respect to their suitability in determining the “Generally Related” relationship and the “Broader/Narrower” relationship.

We hypothesize that the combined similarity measure developed in our work will outperform other existing similarity measures in determining the “Generally Related” relationship. We also hypothesize that considering individual conditional probabilities with unbalanced values in the mutual conditional probability model on the premise of a close “Generally Related” relationship helps to better determine asymmetric relationships (e.g. the “Broader/Narrower” relationship) than symmetric relationships.

In the following experiments, we first choose a relationship determination approach to assign similarity weights to every possible target term pair in a text corpus. A threshold is then set; those relationships having a weight greater than this threshold will be regarded as interesting relationship candidates.

As no relationship determination approach is perfect, the resulting candidate relationships inevitably contain noise, i.e. relationships that are not interesting but

incorrectly weighted highly. The basic principle of our evaluation is to assume that a good relationship determination approach will rank more interesting relationships within a certain number of candidate relationships than a bad approach.

Similar to the automatic evaluation in the last chapter, the interestingness of candidate relationships is judged by their membership in gold standards. The quality of a relationship determination approach is evaluated by the precision and recall value with respect to its ability in including gold standard relationships within a certain number of observed candidate term relationships. Let L denote the total number of relationships in the gold standard; x the number of candidate relationships and y the number of gold standard relationships in x . The precision, recall and F-measure can be defined in a similar way as described in Section 3.3.1.2 in Chapter 3.

Again, we choose astronomy as one of the target domains, and the astronomy thesaurus as the gold standard, which contains two kinds of relationships, the “Broader/Narrower” (bn) relationships – with a total number of 685, and the “Related” (related) relationships – with a total number of 1468. The same web document collection used for evaluating term extraction approaches in the last chapter is used here as a text corpus (cf. Section 3.3.4 for detailed information). For the sake of simplicity and clarity, we only calculate relationships between the gold standard terms in our experiments. There are about 40000 possible pairs of gold standard terms that co-occur in at least three documents in the corpus, among them 743 gold standard relationships (235 bn relationships and 508 related relationships). This set of term pairs is applied for evaluating different approaches regarding their ability to extract gold standard relationships.

We delete all stopwords and other non-topical words and use only the topical terms to build a text segment. We empirically set the length of a text segment to 20, which provides the best performance in explorative tests.

4.6.1 The “Generally Related” Relationship

In this section, we evaluate different approaches with respect to their ability to determine the “Generally Related” relationship. We will therefore not distinguish the bn- and related relationship in the gold standard, and treat them as one general type of relationship.

While it is hardly possible for us to compare all introduced similarity measures for all possible notions of context and commonality of context, previous evaluations [78][80][17] can serve for focusing. Based on these evaluations, we decide to include PMI, COS, and DICE in our evaluation, because PMI has been shown to be the best similarity measure for synonym detection [80][82] and COS and DICE are the two other most popular similarity measures which are not sufficiently evaluated in these evaluations. JAC is equivalent to DICE with respect to ranking. Other similarity measures like occurrence based CHI, YY and content based L1 and CJSD are not evaluated because of their known weak performance for the task at hand [80][78].

Figure 4.6 compares the performance of several variations of the PMI measure for different notions of individual and common context listed in Table 4.4.

Table 4.4. Different notions of context and common context for calculating PMI

	$f(t_i), i=1,2$	$f(t_1, t_2)$
PMI_occ_con	$ \text{context_occ}(t_i) $	$ \text{context_occ_con}(t_1, t_2) $
PMI_occ	$ \text{context_occ}(t_i) $	$ \text{context_occ}(t_1, t_2) $
PMI_con	$ \text{context_con}(t_i) $	$ \text{context_con}(t_1, t_2) $
PMI_occ_doc	$ \text{context_occ_doc}(t_i) $	$ \text{context_occ_doc}(t_1, t_2) $

The size of the text segments is set to 20 terms to the left and to the right of a target term, with the exception of the PMI_occ_doc, where context_occ_doc is a special type of

context_occ by using entire documents as basic text segments. We use PMI_occ_doc as the baseline for comparison.

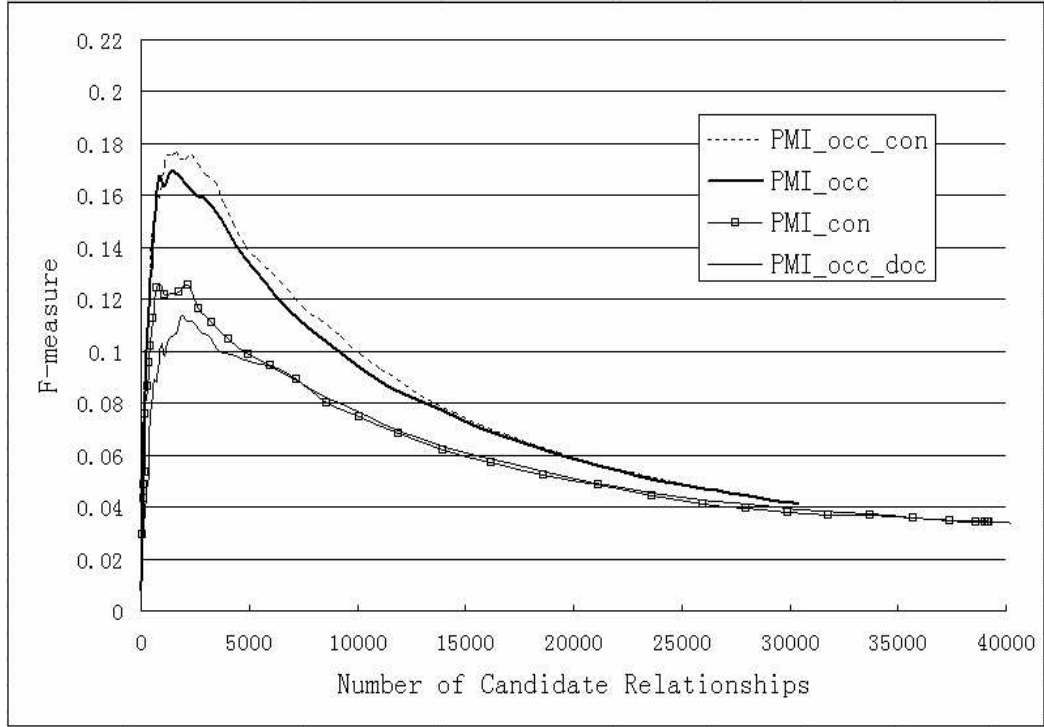


Figure 4.6. Comparison of different variations of PMI

Figure 4.7 compares the performance of several variations of the COS measure with the performance of DICE_occ_freq and our measure rel_sqrt_occ. The COS_freq formula introduced in Section 4.2.2 is used here to calculate COS_occ_freq, COS_con_freq, and COS_occ_con_freq, whereby the context c in COS_freq is instantiated by the corresponding individual context and common context. Similar to PMI_occ_doc, COS_occ_doc_freq is displayed as base line, where a document is used as a text segment. For a comparison with PMI measures, Figure 4.7 is scaled to the same size as Figure 4.6.

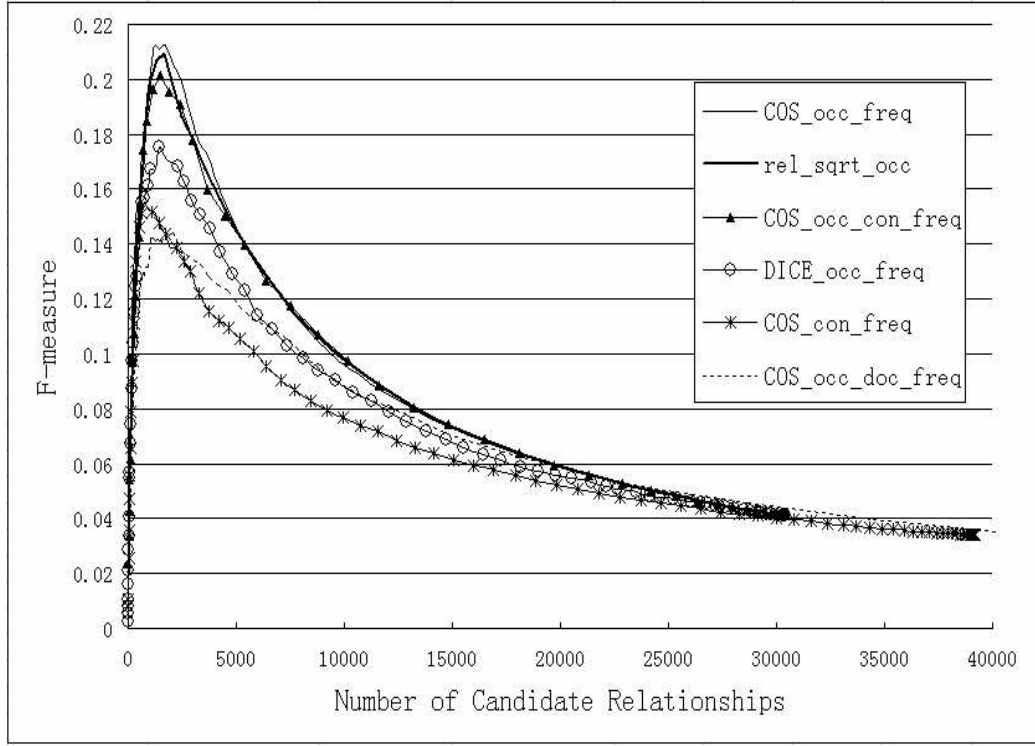


Figure 4.7. Comparison of other similarity measures

The figures clearly show that measures based on occurrence contexts covering entire documents, i.e. PMI_occ_doc and COS_occ_doc_freq, generally perform worse than contexts based on smaller text segments.

As also discussed in Section 4.2.2 and Section 4.3.2, the similarity measures based on pure content context (i.e. PMI_con and COS_con_freq) generally perform significantly worse than those based on occurrence context. In fact, they perform only slightly better than the baselines which use documents as text segments. This result is consistent with the evaluation of Terra et. al. [80], who observed an unexpectedly poor performance of content based approaches for detecting highly related terms.

Among the occurrence based approaches, PMI_occ and DICE_occ_freq perform worse than COS_occ_freq and rel_sqrt_occ. PMI_occ has a maximum F-measure F_{max}

= 0.1687, whereas the F-max of COS_occ_freq is 0.2126, achieving an improvement of almost 26%.

No significant difference can be observed between COS_occ_freq, which takes into account the actual frequencies of terms occurring in a context, and rel_sqrt_occ, which only considers the number of individual and common occurrence contexts of terms.

The measures based on common context combining occurrence and context, i.e. PMI_occ_con and COS_occ_con_freq, do not perform significantly differently from their occurrence based counterparts PMI_occ and COS_occ_freq, with PMI_occ_con slightly better than PMI_occ, and COS_occ_con_freq slightly worse than COS_occ_freq.

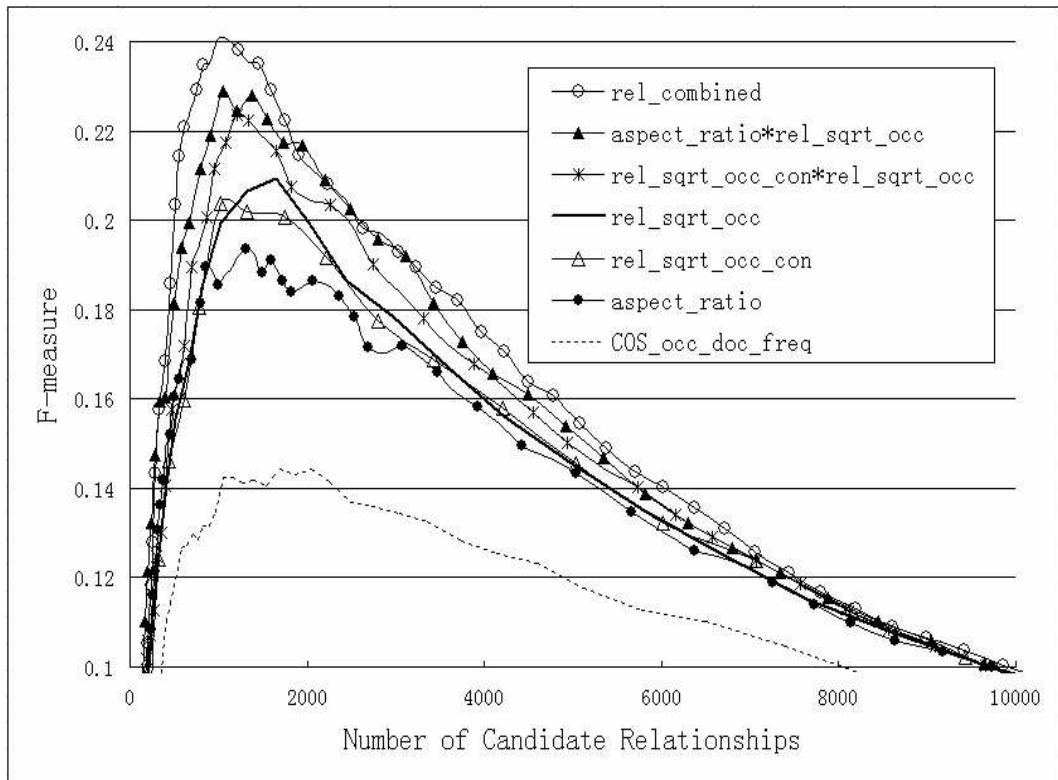


Figure 4.8. Comparison of different similarity measures developed in this work

Figure 4.8 compares the similarity measures introduced in our work and several combinations of them with COS_occ_doc_freq as the baseline. Notice that in this figure,

only the overall trend of a curve is important for analysis. Some measures like the aspect ratio may show small irregularities on their curves due to noise contained in the large amount of data and the different intervals with which the points on the curves are sampled and displayed. Such small irregularities do not affect the final results of comparison, and will be therefore not addressed in the following analysis.

Not much surprisingly all individual measures on the figure perform significantly better than the baseline, with the simple occurrence based measure `rel_sqrt_occ` performing best individually. However, since each of the three measures takes advantage of a different kind of evidence indicating a close relationship, combining these evidence types significantly improves the performance. Figure 4.8 gives two out of the three possible binary combinations `aspect_ratio*rel_sqrt_occ` and `req_sqrt_occ_con*rel_sqrt_occ`; `aspect_ratio*rel_sqrt_occ_con` is not displayed because it performs very similarly to the other binary combinations. Finally, the combination of all three individual measures, `rel_combined`, clearly outperforms all other measures. With its $F\text{-max} = 0.2407$, it improves the performance of the best measure in Figure 4.7 `COS_occ_freq` ($F\text{-max} = 0.2126$) by almost 13%, and the performance of the baseline `COS_coo_doc_freq` ($F\text{-max} = 0.1424$), which is the most frequently applied approach in automatic thesaurus construction [73][40], by almost 70%.

4.6.2 The “Broader/Narrower” Relationship

In this section we automatically evaluate different approaches with respect to their ability to find the “Broader/Narrower” relationship.

As in the previous evaluation, a certain number of candidate relationships is first calculated by applying each method. We then determine the respective number of the two kinds of gold standard relationships in the candidate relationships, i.e. the “Broader/Narrower” (bn) relationship and the “Related” (related) relationship. Let L_{bn} and $L_{related}$ denote the total number of bn and related relationships in the gold standard

respectively; x the number of candidate relationships and y_{bn} and $y_{related}$ the number of the bn and related relationships in the x candidate relationships respectively. The recall values of the bn and related relationships are defined as follows:

Recall of bn relationships: $r_{bn} = y_{bn} / L_{bn}$

Recall of related relationships: $r_{related} = y_{related} / L_{related}$

We then calculate the ratio between the two recall values, i.e. $r_{bn}/r_{related}$, to determine the “relative ability” of an approach in finding the bn relationships. Intuitively, an approach favoring bn relationship will tend to rank relatively more bn relationships in the top x candidate relationships than the related relationships, thereby resulting in a larger value of $r_{bn}/r_{related}$. If we take x as the x axis and $r_{bn}/r_{related}$ as the y axis, we can draw curves in a graph for different relationship determination approaches for an intuitive comparison.

According to this principle, several important relationship determination approaches are evaluated in Figure 4.9. Among them, $P_{occ_doc}(t_1|t_2)$ (the one used in Sanderson’s work [76]) and $P_{con}(t_1|t_2)$ are unbalanced conditional probabilities based on occurrence context and content context, respectively, by using documents as text segments. $rel_combined$, rel_sqrt_occ , aspect ratio, $rel_sqrt_occ_con$ and rel_sqrt_con are mutual conditional probabilities based on different context types.

When calculating the unbalanced conditional probabilities, we always deliberately assign the notation t_1 and t_2 to a pair of terms so that $P(t_1) \geq P(t_2)$, which is equivalent to $P(t_1|t_2) \geq P(t_2|t_1)$. In Figure 4.9, we only consider how the larger conditional probability, i.e. $P(t_1|t_2)$, affects the determination of the bn relationship, with the implication that the other conditional probability, i.e. $P(t_2|t_1)$, being always smaller than or equal to $P(t_1|t_2)$, as shown in the following formulae.

$$\begin{cases} P(t_1 | t_2) \geq th_2 \\ P(t_2 | t_1) \leq P(t_1 | t_2) \end{cases}$$

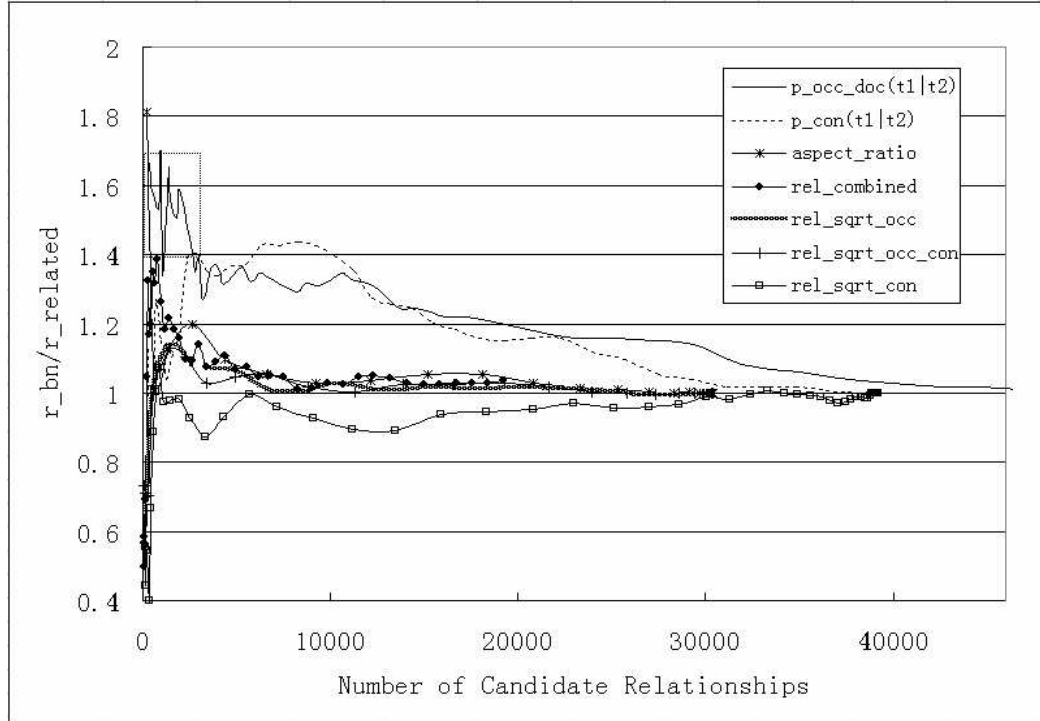


Figure 4.9. Comparing the relative ability of different similarity approaches in finding bn relationships

As we can see in Figure 4.9, almost all approaches based on mutual conditional probability lie very closely to the line $r_{bn}/r_{related} = 1$, which is used as the baseline, meaning that the ability of these approaches to find the bn relationships is almost equal to their ability to find the related relationship. Their relative abilities in finding bn relationships are therefore rather low. Among these approaches, however, we notice that the aspect ratio and rel_combined lie generally slightly higher than others, which confirms our conclusion in 4.4.1 stating that aspect ratio tends to retrieve more relationships between general terms and relatively specific terms, including the “Broader/Narrower” relationships. Since rel_combined integrates aspect ratio as a part of it, it also tends to slightly favour “Broader/Narrower” relationships.

In contrast, the approaches based on unbalanced conditional probabilities, i.e. $P_{occ_doc}(t_1|t_2)$ and $P_{con}(t_1|t_2)$, lie clearly far above the baseline, with $P_{occ_doc}(t_1|t_2)$

providing the best overall performance. For the sake of clarity, we do not draw $P_{occ}(t_1|t_2)$ and $P_{occ_con}(t_1|t_2)$ in the diagram, which lies between $P_{con}(t_1|t_2)$ and aspect ratio.

Let us now focus on the $P_{occ_doc}(t_1|t_2)$. We see that the performance of the approach decreases when the threshold th_2 decreases, resulting in a generally declining curve of $P_{occ_doc}(t_1|t_2)$ which converges to the x-axis. The conditional probability $P_{occ_doc}(t_1|t_2) \geq th_2$ provides the best performance when th_2 takes a value between 0.7 and 0.8 – we mark the corresponding part of the curve with a dotted line square in the graph. It is worth noting that each curve on the figure fluctuates to a great extent at its beginning phase due to the relatively small number of observed relationships. However, we can still clearly see that the curve of $P_{occ_doc}(t_1|t_2)$ fluctuates mainly between 1.4 and 1.8, which is much higher than the fluctuation ranges of other curves that normally lie between 0.4 and 1.4.. As th_2 further decreases, the performance of the approach falls rapidly. After th_2 drops below 0.1, the curve tends to converge to the baseline like other approaches in the graph. These results are consistent with the work of Sanderson [76], where th_2 is empirically set to a constant value of 0.8.

Figure 4.9 shows that within a certain number of top ranked relationships weighted by unbalanced conditional probabilities, relatively more bn relationships can be found than related relationships, which, however, does not imply a large absolute number of bn relationships. In contrast, as mentioned in Section 4.5, since unbalanced conditional probabilities can also be observed with many term pairs having a rather remote relationship, these term pairs will be incorrectly highly ranked, which reduces the number of interesting relationships in the top ranked candidate relationships. The absolute number of interesting bn relationships is therefore very small, although it is much larger than the number of related relationships.

In Figure 4.10 we compare several approaches with respect to their “absolute” ability to find bn relationships. As in previous experiments, we use the number of candidate relationships as the x axis, and the F-measure regarding bn relationships,

denoted as f_{bn} , as the y axis, which combines the precision and recall in retrieving absolute number of bn relationships from the total candidate relationships.

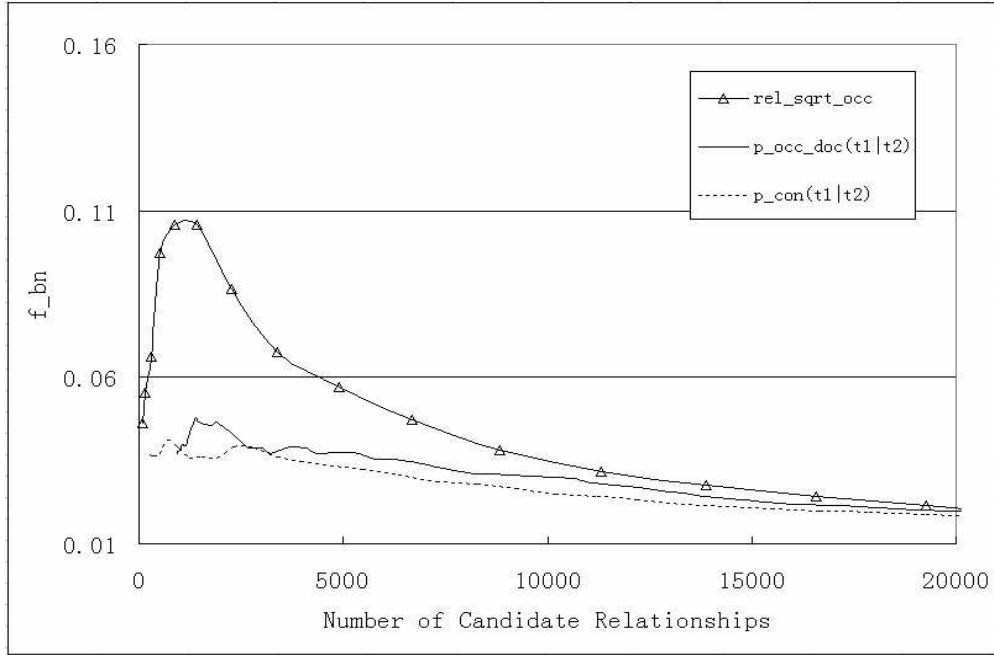


Figure 4.10. Comparing the absolute ability of different similarity measures to find bn relationships

As displayed in the figure, the two conditional probabilities $P_{occ_doc}(t_1|t_2)$ and $P_{con}(t_1|t_2)$, which are shown to have high relative ability in retrieving bn relationships, perform poorly when considering the absolute number of bn relationships as a criterion for evaluation. In contrast, rel_sqrt_occ provides a much better performance, because it finds many more “Generally Related” relationships, which includes both bn and related relationships. Although the last experiment shows that rel_sqrt_occ does not specially favour bn relationships, the absolute number of the bn relationships increases as the number of the “Generally Related” relationships increases – so does the absolute number of the related relationships.

As suggested in Section 4.5, one effective way to address this problem is to first determine a set of close “Generally Related” relationships by using mutual conditional

probability with an appropriate threshold; then apply unbalanced conditional probabilities to this restricted set of relationships to further determine the bn relationships. In this way, term pairs having unbalanced conditional probabilities but rather remote relationships (the number of such term pairs is usually quite large in a corpus) are automatically filtered out, which helps to improve the precision to a great extent.

To enable a direct comparison with the work of Sanderson [76], we use rel_sqrt_occ to instantiate the mutual conditional probability, and P_occ_doc for the conditional probabilities, as shown in the following formulas. The results are displayed in Figure 4.11.

$$\begin{cases} rel_sqrt_occ(t_1, t_2) = P_occ(t_2 | t_1) \times P_occ(t_1 | t_2) \geq th_1 \\ P_occ_doc(t_1 | t_2) \geq th_2 \\ P_occ_doc(t_2 | t_1) \leq P_occ_doc(t_1 | t_2) \end{cases}$$

We distinguish two kinds of curves in the figure: the base lines, including rel_sqrt_occ and P_occ_doc and the thresholding lines. Each of the five thresholding lines corresponds to a fixed value of the threshold th_1 for rel_sqrt_occ , representing a certain level of relatedness with respect to the “Generally Related” relationship. The larger the th_1 , the closer t_1 and t_2 are generally related. Different points in an individual curve correspond to different values of the threshold th_2 for P_occ_doc , which ranges from 1 to 0.

As we can see in the figure, the best performance is provided when th_1 is set to 0.12, which is a rather high relatedness level. At this relatedness level (i.e. with th_1 being fixed at the value 0.12), we try different th_2 values from 1 to 0 with an interval of 0.02 and calculate f_bn for each th_2 value, thereby forming the curve of $rel_sqrt_occ \geq 0.12_p_occ_doc(t_1|t_2)$. It is clear that the curve lies consistently much higher above the baseline curves of rel_sqrt_occ and p_occ_doc , meaning that we can find many more bn relationships by using unbalanced conditional probability on the premise of the mutual conditional probability instead of using them separately. This

thresholding curve reaches its maximum when th_2 is about 0.8, where it also has the largest distance from the curve of rel_sqrt_occ . At this point, $rel_sqrt_occ \geq 0.12_p_occ_doc(t_1|t_2)$ can find almost three times more bn relationships than the rel_sqrt_occ . This is consistent with the previous results shown in Figure 4.9.

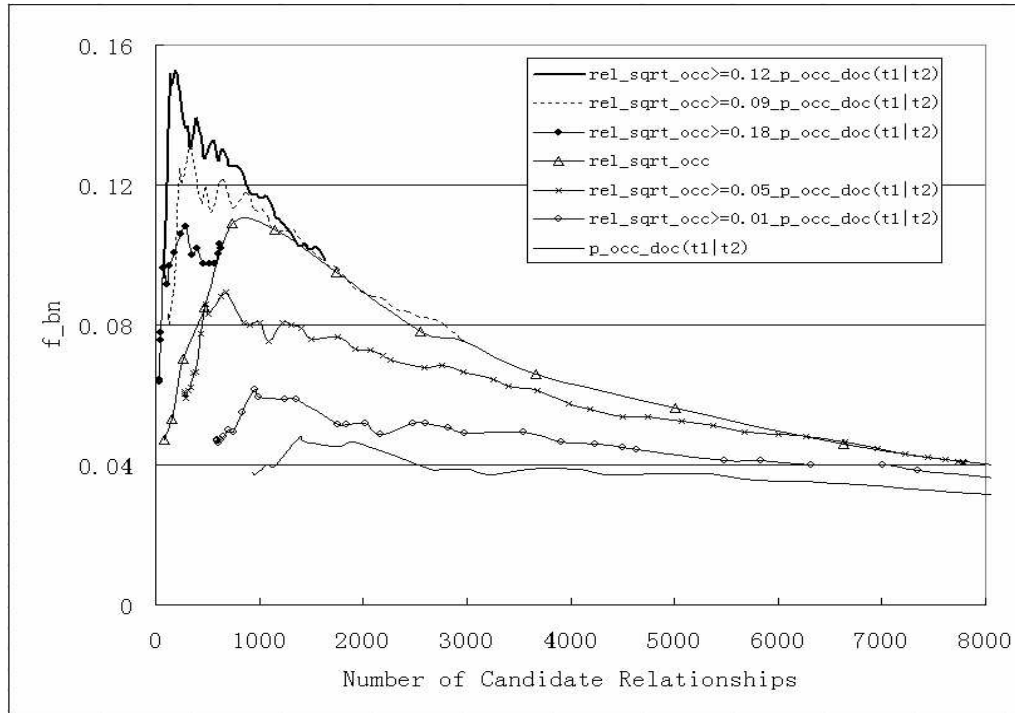


Figure 4.11. Combining rel_sqrt_occ and P_occ_doc

When th_1 decreases, the relatedness level also decreases. More and more noise, i.e. term pairs with unbalanced conditional probabilities but remote relationships, will be included, thereby decreasing the performance of finding the real bn relationships. The $rel_sqrt_occ \geq 0.09_p_occ_doc(t_1|t_2)$ with $th_1=0.09$, for example, lies below the curve of $rel_sqrt_occ \geq 0.12_p_occ_doc(t_1|t_2)$ with $th_1=0.12$. And the curves with lower th_1 values e.g. $rel_sqrt_occ \geq 0.05_p_occ_doc(t_1|t_2)$ and $rel_sqrt_occ \geq 0.01_p_occ_doc(t_1|t_2)$ lie even below rel_sqrt_occ . If th_1 is set to 0, the thresholding curve will completely overlap the curve of P_occ_doc that lies at the bottom of the graph, because there is no restriction

applied any more by the mutual conditional probability. Unbalanced conditional probabilities are used in this case alone for finding the bn relationships.

When th_1 is too large, e.g. $th_1=0.18$, only a very small number of “Generally Related” relationships – thus even fewer bn relationships – can be included, resulting in a rather low value of recall (r_{bn}) and thereby a relatively low value of the F-measure (f_{bn}). That is why the curve of $rel_sqrt_occ \geq 0.18_p_occ_doc(t_1|t_2)$ lies much lower than the $rel_sqrt_occ \geq 0.12_p_occ_doc(t_1|t_2)$. Note that although the maximum point of this curve is achieved with $th_2=0.55$, the maximum distance from the rel_sqrt_occ is still achieved with $th_2=0.8$.

It is worth noting that all thresholding curves end up at the curve rel_sqrt_occ , because at the end point of a threshold curve the value of th_2 is 0, which means that only the mutual conditional probability rel_sqrt_occ is used without considering the unbalanced conditional probability.

In previous experiments, we have used only one conditional probability, i.e. $P(t_1|t_2)$, to find the bn relationships with the implication that $P(t_2|t_1) \leq P(t_1|t_2)$. We now discuss how the thresholding of the smaller conditional probability $P(t_2|t_1)$, i.e. different values of th_3 in the following formulae, will affect the results of experiments.

$$\begin{cases} rel_sqrt_occ(t_1, t_2) = P_occ(t_2 | t_1) \times P_occ(t_1 | t_2) \geq th_1 \\ P_occ_doc(t_1 | t_2) \geq th_2 \\ P_occ_doc(t_2 | t_1) \leq th_3 \end{cases}$$

Depending on the way that we assign the notations t_1 and t_2 , th_3 will always be smaller than or equal to th_2 .

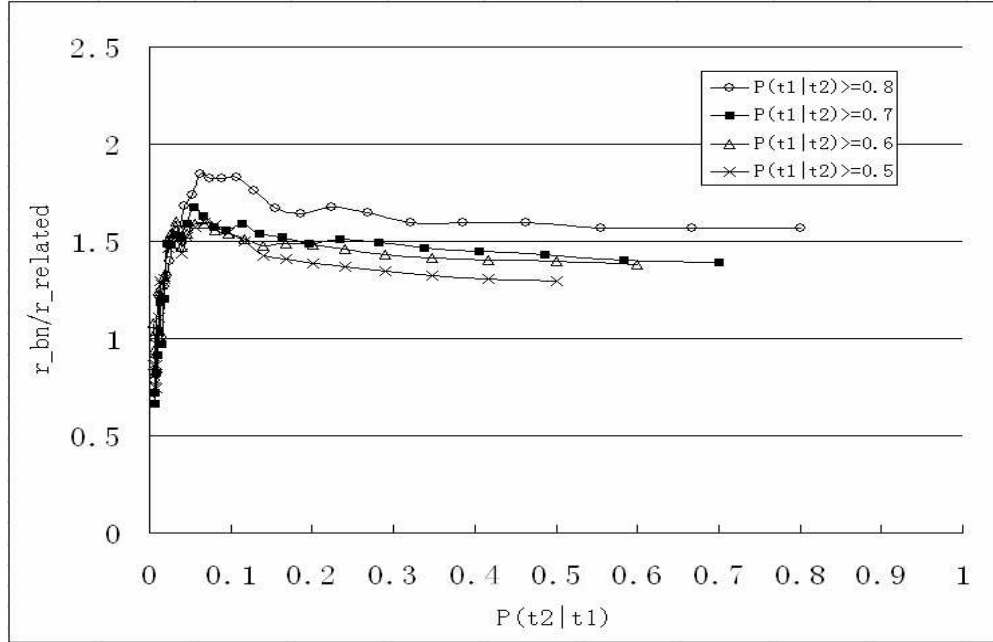


Figure 4.12. Combinations of $P(t_2|t_1)$ and $P(t_1|t_2)$

In Figure 4.12 we use the value of $P(t_2|t_1)$ as the x axis, which ranges from 0 to 1 and the ratio between r_{bn} and $r_{related}$, i.e. $r_{bn}/r_{related}$ as the y axis. Each curve in the graph corresponds to a relatively high th_2 level. Being consistent with previous experiments, the highest curve with $th_2=0.8$ provides the best overall performance in finding the relative number of bn relationships compared to related relationships. The performance diminishes when th_2 takes smaller values such as 0.7, 0.6 and 0.5. If we fix on one individual curve, we can see that the curve remains almost unchanged when th_3 varies between 0.1 and 1. However, there is a slight rise with every curve when th_3 falls between 0.05 and 0.1. The curves fall rapidly as th_3 becomes smaller than 0.05. This result suggests that, after confining th_2 to a larger value, e.g. 0.8, restricting th_3 to a smaller value, e.g. a value between 0.05 and 0.1, will further slightly improve the performance. However, in order to keep the simplicity and a higher performance-cost ratio, one only needs to make restrictions to th_2 and require th_3 to be smaller than or equal to th_2 .

4.6.3 Automatic Evaluation in a Second Domain

To check the consistency of the experimental results achieved in the domain of astronomy, we also carry out another automatic evaluation in the construction domain, which covers both architecture and civil engineering.

The construction thesaurus introduced in Section 3.3.5 is employed again as a gold standard, which contains a total of 21320 relationships, among them 8657 bn relationships and 12663 related relationships. The same web document collection applied to evaluate concept extraction approaches in the construction domain is used again as a text corpus.

In contrast to the astronomy corpus, which tends to contain many rather long and technical documents, the texts contained in the construction corpus are usually quite small (with an average length of 39 terms after deleting domain non-topical terms) and are not especially concerned with construction techniques, but normally concerned with introduction of companies that provide different products and services in the business of construction and maintenance. With a total number of 19154 documents, the construction corpus is much larger than the astronomy corpus, which has only 6876 documents. As in the astronomy domain, we only consider relationships between the gold standard terms for evaluation. We arbitrarily match each gold standard term with every other gold standard term to form term pairs. Those term pairs occurring in at least three documents in the construction corpus are kept, resulting in about 98751 candidate relationships, with 485 of them being real gold standard relationships (230 bn relationships and 255 related relationships). This set of term pairs is applied for testing different approaches with respect to extraction of gold standard relationships. As in the astronomy domain, we delete all stopwords and other non-topical words and use only the topical terms to build a text segment. A value of 20 terms is also here the optimal length of text segment.

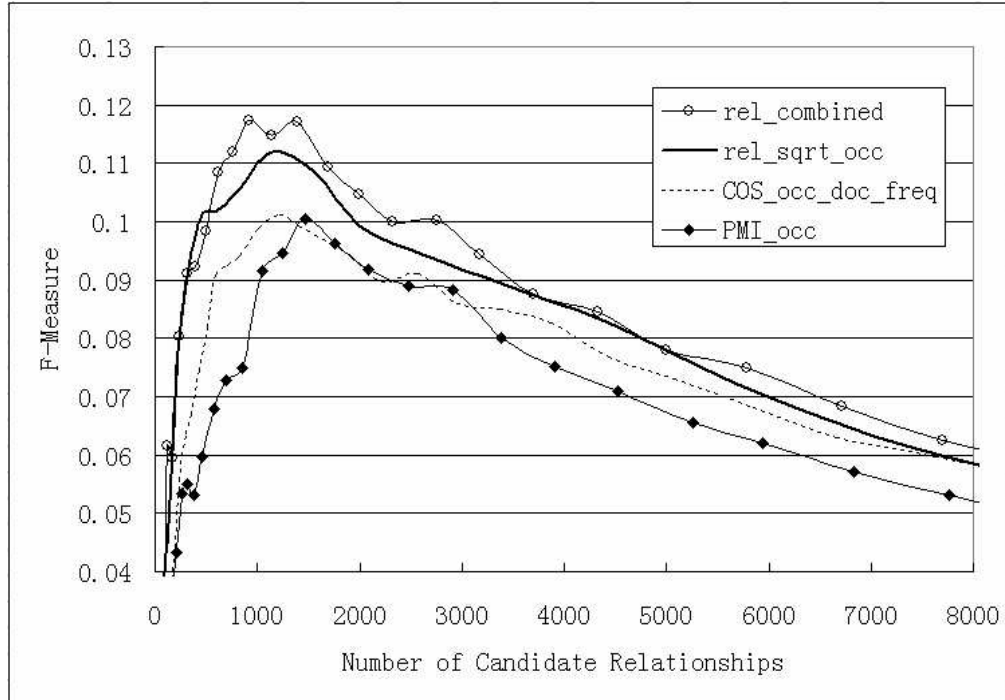


Figure 4.13. Comparison of different similarity measures in a second domain

Figure 4.13 displays the curves of some of the most important similarity measures. The result is consistent with the previous experiments: while the combined similarity measure `rel_combined` still provides the best performance, `rel_sqrt_occ` performs better than `COS_occ_doc_freq` and `PMI_occ`. A noticeable difference here is that in this graph, `COS_occ_doc_freq` lies much closer to `rel_combined` and `rel_sqrt_occ` than in Figure 4.8. In fact, `rel_combined` achieves an F-max improvement of only 16% over `COS_occ_doc_freq`, which is much lower than that in the astronomy experiments with nearly 70% improvement. The reason lies in the fact that, as also mentioned above, the documents in the construction corpus are usually rather small. They have an average length of 39 terms, which is not much larger than the optimal text segment length of 20 terms, suggesting that a document in the construction corpus may contain far fewer topics than a document in the astronomy corpus. Restricting text segments to a smaller window

will therefore not provide an improvement as great as in astronomy domain, where the average document length is 116 terms after deleting domain non-topical terms.

Furthermore, we have checked how the conditional probability $P_{occ_doc}(t_1|t_2)$ helps to better find bn relationships in the construction domain. Figure 4.14 shows similar results as in domain astronomy. The curve of $P_{occ_doc}(t_1|t_2)$ lies much higher above the base line of $r_{bn}/r_{related}=1$, meaning that it is capable of finding more bn relationships than related relationships. The curve has the best performance when th_2 is around 0.8 – the corresponding curve part is marked with a smaller dotted line square. The second best performance is achieved when th_2 is around 0.6 – marked with a larger dotted line square. With this threshold, more candidate relationships can be covered. In contrast to $P_{occ_doc}(t_1|t_2)$, the mutual conditional probability rel_sqrt_occ does not favour bn relationships and therefore almost completely overlaps the baseline.

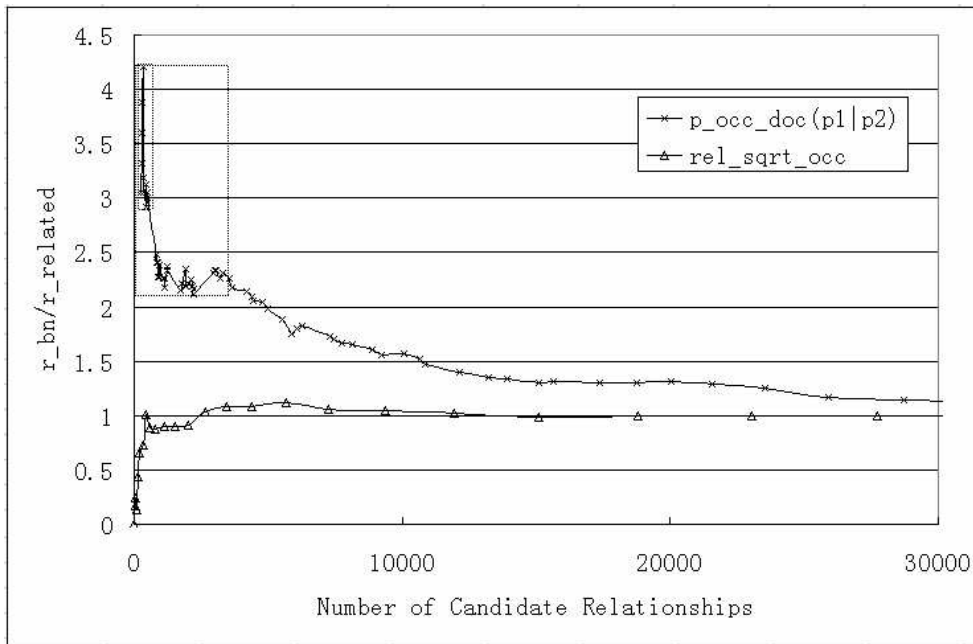


Figure 4.14. Relative abilities of similarity measures in finding bn relationships

Figure 4.15 shows that, also in the construction domain, the combinations of $P_{occ_doc}(t_1|t_2)$ and rel_sqrt_occ with several fixed thresholds provide fairly good

performance with respect to finding bn relationships. With $th1=0.05$ and $th2=0.8$, the curve $rel_sqrt_occ \geq 0.05_p_occ_doc(t1|t2)$ has the largest distance from the curve rel_sqrt_occ .

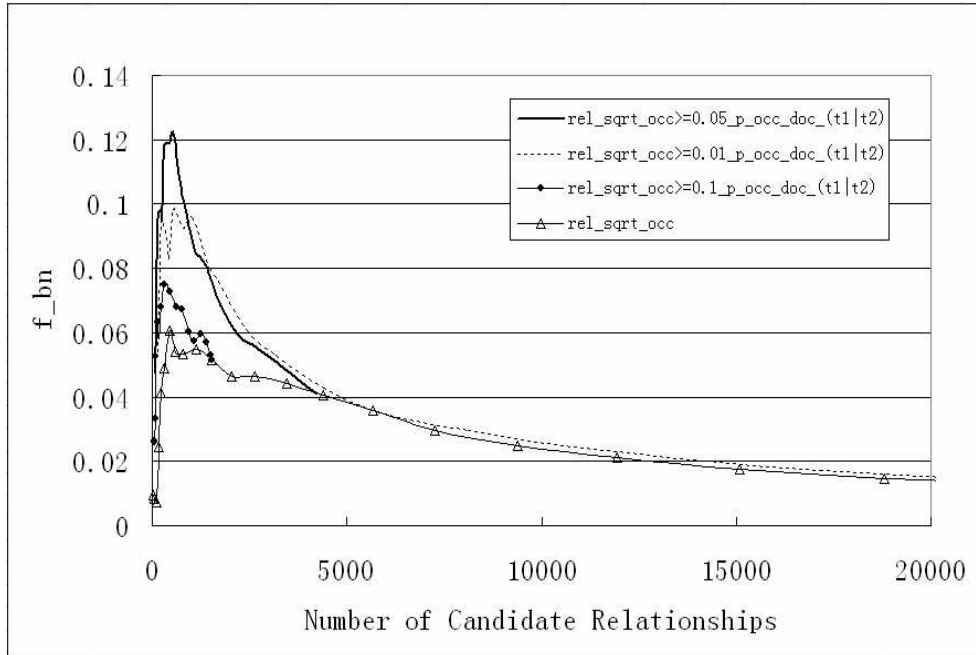


Figure 4.15. Absolute abilities of similarity measures in finding bn relationships

4.7 MANUAL EVALUATION AND EXAMPLES

4.7.1 $rel_sqrt_occ_con$

In the automatic evaluation for the astronomy domain we have seen that $rel_sqrt_occ_con$ provides a good complement to rel_sqrt_occ by considering content of context. Both measures achieve their respective F-max when the number of candidate relationships is around 1200.

We therefore choose the top 1200 candidate relationships ranked by each of the two measures to form two relationship sets and label them as S_1 (for rel_sqrt_occ) and S_2 (for

rel_sqrt_occ_con). Table 4.5 lists some statistics of the two sets, where we use $S_1 \cap S_2$ to denote the intersection of the sets S_1 and S_2 , and $S_i - S_j$ denote the relative complement of S_j in S_i .

Table 4.5. Statistics of S_1 and S_2

	Number of total relationships	Number of gold standard relationships
S_1	1200	208
S_2	1200	198
$S_1 \cap S_2$	561	133
$S_1 - S_2$	639	75
$S_2 - S_1$	639	65

It is clear that both measures perform well for the top 1200 candidate relationships, where rel_sqrt_occ works a little better than rel_sqrt_occ_con, because S_1 contains slightly more gold standard relationships than S_2 (208 vs. 198). The intersection of S_1 and S_2 has only 561 candidate relationships, constituting about 47% of the total relationships in S_1 or S_2 . This means that more than half of the relationships contained in S_1 and S_2 are different, resulting in a relatively large set of $S_1 - S_2$, i.e. the relative complement of S_2 in S_1 , and $S_2 - S_1$, i.e. the relative complement of S_1 in S_2 , both sets with 639 relationships. We then sort these relationships according to the number of documents containing both terms (doc_same) in the relationships in descending order. Table 4.6 lists the top 30 relationships in $S_1 - S_2$ and $S_2 - S_1$ respectively, with the column “Match” showing if the corresponding relationship is contained in gold standard thesaurus. As we can see, although the number of “Matches” is relatively small (because the number of candidate relationships (98751) is much larger than the number of relationships contained in the gold standard (485)), most of the relationships in the table are intuitively interesting.

Table 4.6. Top 30 Relationships in S1-S2 (the relative complement of S2 in S1) and S2-S1 (the relative complement of S1 in S2). The column “Match” indicates if the corresponding relationships are matched in the gold standard, with “M” for Matched and “U” for Unmatched.

S1-S2			
T ₁	T ₂	doc_same	Match
Earth	Sun	1015	U
Earth	space	896	U
Earth	Moon	846	U
Sun	Moon	801	U
astronomy	Stars	795	U
astronomy	space	739	U
astronomy	Planets	726	U
astronomy	Moon	701	U
Earth	Planets	690	M
astronomy	sky	673	U
Moon	Planets	667	U
space	Sun	663	U
Earth	solar system	660	M
Sun	Stars	638	M
Earth	Stars	617	U
Sun	solar system	610	U
astronomy	Universe	594	U
space	Moon	581	U
space	Stars	574	U
space	Planets	557	U
Earth	Universe	551	U
Earth	sky	547	U
Sun	sky	538	U
space	Universe	536	U
space	solar system	535	U
Earth	Images	532	U
astronomy	solar system	531	U
Moon	Stars	515	U
Stars	Universe	502	U
Moon	solar system	486	U

S2-S1			
T ₁	T ₂	doc_same	Match
Astrophotography	CCD	90	U
water	Air	56	U
mathematics	Chemistry	52	U
Chemistry	geology	39	U
eyepieces	Filters	34	U
galaxies	large scale structure	33	U
evolution	dynamics	32	U
equator	northern hemisphere	30	U
ice	water ice	29	M
ecliptic	precession	26	U
Atoms	photons	26	U
CCD	cameras	25	U
Atoms	intensity	25	U
Relativity	quantum mechanics	24	U
magnetosphere	auroras	24	U
black holes	Gamma Ray Bursts	23	U
instruments	Instrumentation	22	M
spectrum	gamma rays	21	U
Relativity	acceleration	19	U
Sunspots	Solar Cycle	19	U
electrons	magnetic fields	19	U
Perihelion	inclination	18	U
Relativity	quantum theory	17	U
particles	Neutrinos	17	M
seasons	summer solstice	17	U
variable stars	Double Stars	17	U
particles	charged particles	16	U
spectra	absorption lines	16	M
cameras	lenses	16	U
Pulsars	Gamma Ray Bursts	16	U

Table 4.6 shows that the term pairs contained in S1-S2 are usually quite general in the domain of astronomy, with large values of doc_same, while term pairs contained in

S2-S1 are relatively specific, with much smaller values of doc_same. In fact, the average value of doc_same in the whole set of S1-S2 is 88.17, and that in S2-S1 is only 6.13.

These examples clearly reflect the fact that rel_sqrt_occ and rel_sqrt_occ_con consider different statistical evidence in text corpora for relationship determination. While both approaches consider the co-occurrence information as the most important evidence, rel_sqrt_occ stresses the number of text segments containing the co-occurrences without further considering the content of the text segments, and rel_sqrt_occ_con focuses on the variety of context, i.e. the number of unique content terms in the context of co-occurrences, where the number of text segments does not play an important role.

The major disadvantage of rel_sqrt_occ is that it tends to weight relationships too strongly when the contexts of co-occurrence have similar content. One example is “earth” and “stars”. They are two rather general terms occurring in many text segments and their relationship is thus highly weighted by rel_sqrt_occ with a rank of 357 – within the top 1200 candidate relationships. However, a closer look into the text segments reveals that a lot of these texts are concerned with introductory knowledge of the astronomy domain and tend to have similar content. Additionally, 5 of the 617 web documents containing the two terms have almost identical content to other documents – they are obviously just copies of the same web pages under different URLs. All of these lead to a too strong relationship weighting between “earth” and “stars”. On the contrary, the measure rel_sqrt_occ_sqrt, which takes into account the unique content of contexts, weights the relationship more correctly with a rank of 2377 – outside the top 1200 candidate relationships.

On the other hand, rel_sqrt_occ_con also has a drawback. It tends to give too strong relationship weighting when two target terms co-occur in a few text segments with quite dissimilar contents. We take the term pair “lunar eclipse” and “quadrants” as an example. These two terms closely co-occur in only two documents. The first document is

concerned with “medieval England astronomy”, where the two terms belong to different topics that happen to be located closely to each other. One topic is about the church’s control of astronomical predictions such as lunar eclipse, while the other topic concerns medieval astrological instruments including quadrants. The context terms here could be “church”, “control”, “predictions”, “instruments”, “astrolabes” etc.. The relevant part of the second document is concerned with a letter written in the year of 1635, in which an astronomer named Peiresc asked the missionary priest Agathange de Vendome in Cairo to carry telescopes and quadrants to the top of a pyramid to observe a lunar eclipse. The context terms here could be “Peiresc”, “priest”, “Agathange de Vendome”, “telescopes”, “pyramid” etc.. It is clear that the context terms for the two target terms “lunar eclipse” and “quadrants” in the according text segments are quite dissimilar and their relationship is therefore highly weighted by `rel_sqrt_occ_con` with a rank of 667 – within the top 1200 candidate relationships. The problem here is that the fact that quadrants as an ancient astronomical instrument were used to observe lunar eclipses in the Middle Ages is only a trivial aspect for both “quadrants” and “lunar eclipse”. This is well reflected by the fact that there exist only very few documents where the two target terms closely co-occur, which, however, cannot be effectively dealt with by `rel_sqrt_occ_con`. In contrast, the measure `rel_sqrt_occ` works much better in this case in that it takes the number of the text segments into consideration and weights the relationships correctly with a much lower rank of 6408.

As we know from the automatic evaluation, the combination of `rel_sqrt_occ` and `rel_sqrt_occ_con`, i.e. `rel_sqrt_occ×rel_sqrt_occ_con`, is capable of achieving a better performance for relationship determination than the individual measures, because a large value of `rel_sqrt_occ×rel_sqrt_occ_con` usually requires relatively large values of both `rel_sqrt_occ` and `rel_sqrt_occ_con`, which actually requires that both kinds of evidence, i.e. the number of text segments and the number of unique content terms, are sufficiently considered.

4.7.2 Aspect Ratio

Besides `rel_sqrt_occ_con`, the automatic evaluation shows that the aspect ratio can be also used as a good complement to `rel_sqrt_occ`. For a manual verification we also choose the top 1200 candidate relationships weighted by aspect ratio and form another set of relationships, labeled as S_3 . Note that S_1 and S_2 still denote the sets of the top 1200 candidate relationships for `rel_sqrt_occ` and `rel_sqrt_occ_con` respectively. Table 4.7 lists some statistics of S_3 , compared with S_1 and S_2 .

Table 4.7. Statistics of S_3 compared with S_1 and S_2

	Number of total relationships	Number of gold standard relationships
S_3	1200	186
S_3-S_1	686	62
S_3-S_2	695	72
$S_3-S_1-S_2$	538	34

This table shows that aspect ratio performs a little worse than `rel_sqrt_occ` and `rel_sqrt_occ_con` when used alone, because S_3 contains slightly fewer gold standard relationships than S_1 and S_2 (186 vs. 208 and 198). However, it is shown to provide many more different relationships in the top 1200 relationships than the other two measures. In fact, around 57% of relationships in S_3 are not contained in S_1 (686/1200), around 58% of S_3 relationships are not contained in S_2 (695/1200) and around 45% of S_3 relationships are contained neither in S_1 nor in S_2 . Table 4.8 lists the top 30 relationships in $S_3-S_2-S_1$, sorted by `doc_same` in descending order and with the column “Match” showing the membership of the relationship in the gold standard thesaurus.

Table 4.8. Top 30 relationships in S3-S2-S1. The column “Match” indicates if the corresponding relationships are matched in the gold standard, with “M” for Matched and “U” for Unmatched.

t₁	t₂	doc_same	Match
Stars	Zodiac	70	U
Earth	debris	65	U
Earth	upper atmosphere	51	U
Earth	gases	49	U
Universe	High Energy Astrophysics	48	U
Sun	precession	47	U
Earth	asteroid belt	43	U
Earth	meteoroids	42	U
Sun	eccentricity	41	U
Earth	Orbital Velocity	37	U
Earth	Earth orbit	36	U
Earth	albedo	35	U
Earth	extinction	35	U
Sun	equinoxes	33	U
Earth	outer planets	32	U
solar system	outer planets	32	U
Earth	perigee	27	M
Sun	Solar Cycle	27	U
Stars	binary stars	27	U
Earth	aphelion	26	U
Sun	solstices	26	U
Earth	inner planets	25	U
Sun	aphelion	25	U
Earth	apogee	24	M
Earth	magnetic Poles	24	U
Sun	Massive Stars	24	U
Sun	inner planets	24	U
Sun	perigee	24	U
Earth	Terrestrial Planets	22	U
Moon	apogee	22	M

A noticeable characteristic of the relationships in this table is that each of the relationships is between a rather general term and a relatively specific term, like “stars – Zodiac” and “earth – perigee”. Compared to S3-S1-S2, most of the relationships in S1-S2 are between two general terms with rather large values of doc_same, like “earth – sun”

and “astronomy – stars”, and the relationships in S2-S1 are usually between two relatively specific terms with smaller values of doc_same, like “ecliptic – precession” and “equator – northern hemisphere”.

The results above show that, while the aspect ratio can be used as a stand alone measure with fairly good performance, it is also capable of highly ranking some interesting relationships that cannot be effectively determined by rel_sqrt_occ and rel_sqrt_occ_con. The term pair “earth – perigee” for example, which has a rank of 326 by the aspect ratio, has a much lower rank of 5283 by rel_sqrt_occ and 2332 by rel_sqrt_occ_con. Therefore, a measure combining the three measures, i.e. rel_combined, will achieve a better overall performance for relationship determination, in that more interesting and mutually complementary relationships are ranked at the top.

However, the fact that the set S3-S1-S2 contains many more relationships between general and specific terms than S1-S2 and S2-S1 does not mean that aspect ratio is capable of finding much more broader/narrower relationships than rel_sqrt_occ and rel_sqrt_occ_con. The reason is that besides these relationships, aspect ratio also determines many relationships between terms that are both general and between terms that are both specific. These relationships usually belong to the intersection of S3 and S1 (514 relationships, around 43% of the total relationships) and the intersection of S3 and S2 (505 relationships, around 42% of the total relationships). For example, among a total of about 40000 candidate relationships, the relationship “earth – sun” which is between two general terms is highly weighted by aspect ratio with a rank of 378, and “apogee – perigee” is a relationship between two relatively specific terms, but highly weighted by aspect ratio with a rank of 38. Another reason is that many relationships contained in S3-S1-S2 are not necessarily asymmetric. A lot of them are symmetric relationships between general and relatively specific terms. Let us take the term pair “earth – perigee” as an example, although “earth” is very general and “perigee” is relatively specific, it is still a symmetric “Related” relationship. This observation confirms our conclusion in Section

4.6.2 that aspect ratio’s ability in finding the “Broader/Narrower” relationships is slightly better than rel_sqrt_occ , but far outperformed by using unbalanced conditional probabilities.

4.7.3 The “Broader/Narrower” Relationship

In this section, we manually check the performance of using unbalanced conditional probabilities to find the “Broader/Narrower” relationships, when combined with the mutual conditional probability. We choose the top 60 relationships weighted by the following formulae and list them in Table 4.9.

$$\begin{cases} rel_sqrt_occ(t_1, t_2) \geq th_1 \\ P_occ_doc(t_1 | t_2) \geq th_2 \\ P_occ_doc(t_2 | t_1) \leq P_occ_doc(t_1 | t_2) \end{cases}$$

where th_1 is set to 0.327 to limit the number of relationships to 60 and th_2 is set to 0.8, which is shown to be the best threshold for P_occ_doc in the automatic evaluation. The relationships are sorted by $P_occ_doc(t_1|t_2)$ in descending order.

Consistent with the results in the automatic evaluation, we see 11 “BN” relationships and 7 “Related” relationships in the table. Since the total number of BN relationships in the gold standard is less than half of the number of Related relationships (235 vs. 508), the ratio of their recall values is relatively high ($(11/235) / (7/508) = 3.4$). We notice that there are also some obvious “BN” relationships that are not included in the gold standard, e.g. “Universe – Inflationary Universe”, “dark matter – baryonic dark matter” etc.

Table 4.9. Top 60 Relationships weighted by combination of conditional probabilities. The column “Match” indicates if the corresponding relationships are matched in the gold standard, with “M” for Matched, “U” for Unmatched and “BN” for “Broader/Narrower” relationship.

T ₁	T ₂	Match
Universe	Inflationary Universe	U
light	past light cone	U
galaxies	superclusters	U
cosmology	Anthropic Principle	R
black holes	Hawking radiation	R
dust	interstellar reddening	U
temperature	CNO Cycle	U
optics	fiber optics	BN
rotation	solar day	U
meteorites	micrometeorites	BN
meteorites	IRON METEORITES	BN
meteorites	ACHONDRITES	BN
Quasars	extragalactic radio sources	BN
Quasars	Seyfert galaxies	U
dark matter	axions	U
dark matter	baryonic dark matter	U
Neutron Stars	bursters	U
X rays	X ray emission	U
Interstellar Medium	Interstellar Molecules	U
molecules	molecular astrophysics	U
luminosity	Hydrogen Burning	U
Celestial Equator	hour angle	U
Dark Nebulae	absorption nebulae	R
interstellar dust	interstellar reddening	U
surface gravity	Effective temperature	U
Hubble constant	Hubble Radius	U
reddening	interstellar reddening	BN
comets	Long period comets	BN
ethane	ethylene	U
galactic rotation	Galactic Poles	U

T ₁	T ₂	Match
Roche lobe	Mass Transfer	U
Universe	Big Bang theory	U
meteorites	Tektites	BN
Universe	Expanding Universe	U
solar system	Oort cloud	U
quarks	leptons	R
galaxy	spiral arms	U
Sun	ecliptic	U
Night Vision	range finders	U
Sun	Oort cloud	U
spectrum	hydrogen lines	U
black holes	white holes	R
Universe	fluctuations	U
meteorites	CHONDRITES	BN
photons	gravitons	U
eccentricity	Semimajor axis	U
event horizon	Naked Singularity	U
umbra	penumbra	R
SETI	ExtraTerrestrial Intelligence	U
astronomy	archaeoastronomy	BN
AGN	Seyfert galaxies	U
Sun	solar wind	U
Perihelion	aphelion	R
electrons	protons	U
telescopes	eyepieces	U
molecules	Interstellar Molecules	BN
prominences	shadow bands	U
equilibrium	Radiative transfer	U
Schwarzschild Radius	Schwarzschild metric	U
Hertzsprung Russell Diagram	hydrostatic equilibrium	U

A closer look at the table further reveals that using unbalanced conditional probabilities is not only capable of better finding the “BN” relationships, it generally helps to better retrieve all kinds of asymmetric relationships, where one term in the relationship is more dependent on another, and the other term is less dependent on the first one. These asymmetric relationships include for example the “part of” relationship, such as the “telescopes – eyepieces”, “galaxy – spiral arms”; the “property of” relationship, such as “sun – ecliptic” (ecliptic is the “path” of the sun in the sky), “Roche Lobe – Mass Transfer”; the “cause” relationship, such as “Neutron Stars – bursters”, “Sun – Solar wind” and “Interstellar dust – Interstellar rendering”.

4.8 CONCLUSION

This chapter has dealt with statistical relationship determination among terms, which is one of the key issues in automatic construction of concept structures. We have presented a mutual conditional probability model as a general framework for determining a “General Relatedly” relationship. We have provided a systematic analysis of two kinds of important approaches – occurrence based approaches and content based approaches. Based on the model and the analysis, a new type of common context and a combined similarity measure combining different kinds of probabilistic evidence have been proposed for better relationship determination. We have also shown that using unbalanced conditional probabilities on the premise of a relatively large mutual conditional probability helps improve finding the “Broader/Narrower” relationships. Experimental results of both automatic and manual evaluation have confirmed our analysis.

CHAPTER 5 CONCLUSION

5.1 SUMMARY OF THE THESIS

Domain-specific concept structures are a powerful tool to deal with the vocabulary problem, especially the vocabulary mismatch problem, which is one of the greatest challenges in information processing tasks.

This dissertation focuses on corpus based statistical approaches for automatic construction of domain-specific concept structures by using text corpora as sources and applying purely statistical methods for the automatic processing. The main contributions of this thesis lie in two aspects: concept extraction and relationship determination.

For the task of concept extraction, we have introduced a notion of topicality to indicate the importance of a term in a target domain, which can be further divided into term representativeness and term specificity. A novel approach is developed for calculating term specificity more accurately. Besides the target domain, the approach also distinguishes different reference domains in the whole text collection, and applies a statistical measure called the Distribution Grade to compare the distributions of a term in different domains. By combining representativeness and specificity, we are able to give topicality weighting to each term in the text corpus according to their importance for the target domain. Experimental results have shown that our methods achieve the best performance for concept extraction for a wide range of candidate terms on different kinds of data basis.

For relationship determination, we have presented a mutual conditional probability model, which can serve as a general framework for formalizing the most successful similarity measures and be used to determine a “Generally Related” relationship. Moreover, we have introduced and discussed several notions of context and common

context based on their underlying probabilistic assumptions, and quantified each by means of conditional probabilities. Because all notions have individual strengths and weaknesses, we have suggested a similarity measure that conjunctively combines the evidence provided by each notion of context. Further, we have shown that using unbalanced conditional probabilities on the premise of a relatively large mutual conditional probability helps to better find the “Broader/Narrower” relationships. Experimental results have confirmed our analysis and shown that the combined similarity measure provides the best performance in relationship determination, achieving an improvement of nearly 70% at the best F-measure value compared with the traditional document based co-occurrence analysis.

In addition, we have also demonstrated that web directory systems like Yahoo! and Google Directories provide sufficient domain coverage, and have high flexibility and accessibility. They can therefore serve as a good source for the task of automatic construction of domain-specific concept structures.

For an effective evaluation, we have developed an automatic method, which uses gold standards and F-measure to automatically evaluate the quality of text sources and compare the performance of different approaches for concept extraction and relationship determination.

The automatically constructed concept structures find their applications in many fields, such as supporting manual construction of concept structures in knowledge engineering, query expansion (both automatic and interactive) in Information Retrieval, and document classification.

5.2 FUTURE WORK

As a source for concept extraction, we have only used the home page of each web site in the web directory systems. Experimental results have shown that this rather small

set of web documents is capable of covering a large set of topical concepts in a domain. As the next step, we plan to extend the data basis by collecting more web pages as text sources. We will further follow the meaningful links in each home page, by using more complex web page analysis methods as described in [15]. We expect to extract more topical concepts from this extended data basis.

For the purpose of evaluation, we have automatically compared our results with gold standards in this thesis. In future research, we intend to experiment on more domains with other gold standards to check the consistence of the results. In addition, it would also be interesting to evaluate the concept structures in different applications, such as supporting manual construction of concept structures, query expansion and document classification, whereby we can check how the performance of the whole application systems could be improved when they integrate the concept structures. To this end, we may need to design a proper user interface for the concept structures to facilitate user access, and carry out user studies to evaluate system performance.

In this work, we have focused on purely statistical methods for building domain-specific concept structures. We also plan in the future to consider linguistic evidences in text corpora for the tasks of concept extraction and relationship determination, which requires profound linguistic knowledge and well performing NLP tools. Compared to purely statistical methods, linguistic approaches usually perform more precisely, but are also much more costly and more dependent on the underlying language. We therefore intend to find a way that combines both statistical and linguistic methods to achieve an optimal performance for automatic construction of domain-specific concept structures.

REFERENCES

- [1] Anick, P.; Tiperneni, S.: "The paraphrase Search Assistant: Terminological Feedback for Iterative Information Seeking". In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Berkeley, CA, USA. 1999, 153-159.
- [2] Banko, M.; Brill, E.: "Mitigating the paucity-of-data problem: Exploring the effect of training corpus size on classifier performance for natural language processing", in Proceedings of the 1st International conference on Human Language Technology Research, San Francisco, California. Morgan Kaufmann, 2001.
- [3] Banko, M.; Brill, E.: "Scaling to very very large corpora for natural language disambiguation", ACL2001.
- [4] Baroni, M.; Bisi, S.: "Using cooccurrence statistics and the web to discover synonyms in a technical language". In proceedings of LREC 2004, Lisbon: ELDA.1725-1728.
- [5] Beitzel, S.; Jensen, E.; Chowdhury, A.; Grossman, D.; Frieder, O.: "Using Manually-Built Web Directories for Automatic Evaluation of Known-Item Retrieval", SIGIR 2003, Toronto, Canada.
- [6] Berners-Lee, T.; Hendler, J.; Lassila, O.: The semantic web – a new form of web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American, 2001.
- [7] Blair, D.C.: "Indeterminacy in the subject access to documents". Information Processing and Management, 22, 229-241, 1986.
- [8] Brewster, C.; Alani, H.; Dasmahapatra, S.; Wilks, Y.: "Data Driven Ontology Evaluation". International Conference on Language Resources and Evaluation, Lisbon, Portugal. 2004
- [9] The Brown corpus
http://clwww.essex.ac.uk/w3c/corpus_ling/content/corpora/list/private/brown/brown.html
- [10] Caraballo, S.A.; Charniak, E.: "Determining the Specificity of Nouns from Text". In Proceedings of EMNLP'99, pages 63-70., 1999.
- [11] Caraballo, S.A.: "Automatic construction of a hypernym-labeled noun hierarchy from text", ACL 1999, pages 120-126.

- [12] Chen, H.; Dumais, S. T.: "Bringing order to the web: Automatically Categorizing Search Results", CHI 2000.
- [13] Chen, H.: "Automatic Construction of Networks of Concepts Characterizing Document Databases", IEEE Transactions on Systems, Man and Cybernetics, 22(5):885--902, September/October 1992.
- [14] Chen, H.: "Automatic Thesaurus Generation for an Electronic Community System", Journal of the American society for information science, 1995.
- [15] Chen, J.L.; Zhou, B. Y.; Shi, J.; Zhang, H. J.; Wu, Q. F.: Function-based Object Model Towards Website Adaptation, In Proc. of WWW10, pp. 587-596, May 2001.
- [16] Chen, Z.; Liu, S.; Liu, W.; Pu, G.; Ma, W. Y.: "Building a Web Thesaurus from Web Link Structure", SIGIR 2003.
- [17] Chung, Y.M.; Lee, J.Y.: "A Corpus-Based Approach to Comparative Evaluation of Statistical Term Association Measures", Young Mee Chung and Jae Yun Lee Published online 17 January 2001 Journal of the American Society for Information Science and technology.
- [18] Church, K.W.; Hanks, P.: "Word association norms, mutual information, and lexicography". Computational Linguistics, 16(1):22-29, 1990.
- [19] Cimiano, P.; Staab, S.; Tane, J.: "Deriving Concept Hierarchies from Text by Smooth Formal Concept Analysis", in Proceedings of the GI Workshop "Lehren - Lernen - Wissen - Adaptivität" (LLWA), Fachgruppe Maschinelles Lernen, Wissenentdeckung, Data Mining, Karlsruhe, Germany, 2003.
- [20] Cimiano, P.; Staab, S.; Tane, J.: "Automatic Acquisition of Taxonomies from Text: FCA meets NLP", In: Proceedings of the ECML/PKDD Workshop on Adaptive Text Extraction and Mining, Cavtat--Dubrovnik, Croatia, 2003.
- [21] Cimiano, P.; Pivk, A.; Schmidt-Thieme, L.; Staab, S.: "Learning Taxonomic Relations from Heterogeneous Evidence", ECAI 2004 Workshop on ontology learning and population, 2004
- [22] Cover, T.M.; Thomas, J.A.: "Elements of Information Theory". Wiley-Interscience, New York, 1991.
- [23] Cronen-Townsend, S.; Croft, W.B.: "Quantifying query ambiguity". In Proceedings of HLT, pp. 94-98, 2002.

- [24] Croft, W.B.; Das, R.: "Experiments with Query Acquisition and Use in Document Retrieval Systems". In J. Vidick, editor, Proceedings of the 13th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 349--368. ACM Press, 1990.
- [25] Crouch, C.J.; Yang, B.: "Experiments in automatic statistical thesaurus construction", in Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Copenhagen, Denmark, June 21-24, 1992.
- [26] Curran, J.R.; Moens, M.: "Improvements in automatic thesaurus extraction". In Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition, pages 59-66, 2002.
- [27] Curran, J.R.: "Ensemble Methods for Automatic Thesaurus Extraction". Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 222 - 229, Philadelphia PA, USA. 2002.
- [28] Dagan, I.; Lee, L.; Pereira, F.C.N.: "Similarity-based models of word co-occurrence probabilities". Machine Learning, 34(1-3):43-69, 1999.
- [29] Delcourt, C.: "About the statistical analysis of co-occurrence". Computers and the Humanities, 26, 21-29, 1992.
- [30] Ekmekcioglu, F.C.: "Effectiveness of query expansion in ranked-output document retrieval systems." Journal of Information Science, Vol. 18, pp. 139--147, 1992.
- [31] Faure, D.; Nedellec, C.: "A Corpus based Conceptual Clustering Method for Verb Frames and Ontology Acquisition". In LREC workshop on Adapting lexical and corpus resources to sublanguages and applications, Granada, Spain, 1998.
- [32] Frakes, B.; Baeza-Yates, R.: "Information retrieval - Data Structures & Algorithms", Prentice Hall, Englewood Cliffs, NJ, 1992.
- [33] Frantzi, K.; Ananiadou, S.: "Automatic Term Recognition using Contextual Cues", in Proceedings of Mulsaic 97, IJCAI, Japan, 1997.
- [34] French, R.M.; Labiouse, C.: "Four problems with extracting human semantics from large text corpora". Proceedings of the 24th Annual Conference of the Cognitive Science Society, NJ, 2002.
- [35] Furnas, G.W.; Landauer, T.K.; Gomez, L.M.; Dumais, S.T.: "The vocabulary problem in human-system communication". Communications of the ACM, 30, 964-971, 1987.

- [36] Gauch, S.; Wang, J.: "A corpus analysis approach for automatic query expansion". Proceedings of the Sixth International Conference on Information and Knowledge Management, pp. 278-284, 1997.
- [37] Gilchrist, A.: "Thesauri, taxonomies and ontologies – an etymological note". Journal of Documentation, Volume 59, Issue 1, pp. 7-18, 2003.
- [38] Glover, E.; Pennock, D.M.; Lawrence, S.; Krovetz, R.: "Inferring hierarchical descriptions". In Proceedings of the 11th International Conference on Information and Knowledge Management (CIKM'02), pages 507--514, 2002.
- [39] Grefenstette, G.: "Use of syntactic context to produce term association lists for text retrieval". In Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 89-97. ACM Press, 1992.
- [40] Grefenstette, G.: "Explorations in Automatic Thesaurus Discovery", Kluwer Academic Publishers, USA, 1994.
- [41] Grefenstette, G.: "The World Wide Web as a resource for example-based machine translation tasks". In Proceedings of the ASLIB Conference on Translating and the Computer, volume 21, London, 1999.
- [42] Gruber, T.R.: "A translation approach to portable ontologies". Knowledge Acquisition, 5:199-220, 1993.
- [43] Hearst, M.: "Automatic Acquisition of Hyponyms from Large Text Corpora". In Proceedings of the Fourteenth International Conference on Computational Linguistics, pages 539--545, Nantes, France, July 1992.
- [44] Hindle, D.: "Noun Classification from predicate-argument structures" ACL90, pp. 268-275, 1990.
- [45] ISO 2788-1986 Documentation – Guidelines for the establishment and development of monolingual thesauri.
<http://www.collectionscanada.ca/iso/tc46sc9/standard/2788e.htm>
- [46] Jing, Y. F.; Croft, W.B.: "An Association Thesaurus for Information Retrieval", In RIAO 94 Conference Proceedings, p. 146-160, New York, Oct. 1994.
- [47] Keller, F.; Lapata, M.; Ourioupina, O.: "Using the web to Overcome Data Sparseness". In Proceedings of EMNLP-02, pp. 230--237, 2002.

- [48] Kita, K.; Kato, Y.; Omoto, T.; Yano, Y.: "A Comparative Study of Automatic Extraction of Collocations from Corpora: Mutual Information vs. Cost Criteria". In *Journal of Natural Language Processing*, 1:21-33. 1994.
- [49] Koller, D.; Sahami, M.: "Hierarchically classifying documents using very few words", *Proc. Of the 14th International Conference on Machine Learning*, pp. 170-178, 1997.
- [50] Lawrie, D.J.: "Language Models for Hierarchical Summarization", PHD Thesis, 2003, <http://www-ciir.cs.umass.edu/~lawrie/papers/lawrieThesis.pdf>
- [51] Lawrie, D.; Croft, W.B.: "Discovering and Comparing Topic Hierarchies". In *Proceedings of RIAO 2000*.
- [52] Lewis, D.D.; Croft, W.B.: "Term Clustering of Syntactic Phrases", *Proc. of the Thirteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 385-404, 1990.
- [53] LDC The corpus catalog of the Linguistic Data Consortium
<http://www.ldc.upenn.edu/Catalog/>
- [54] Lin, D.: "Automatic retrieval and clustering of similar words". In *COLING-ACL*, pp. 768-774, 1998.
- [55] Lin, D.; Pantel, P.: "Concept Discovery from Text". In *Proceedings of Conference on Computational Linguistics 2002*. pp. 577-583. Taipei, Taiwan, 2002.
- [56] Liu, B.; Chin, C.; Ng, H.: "Mining topic-specific concepts and definitions on the web", in *Proceedings of the Twelfth International World Wide Web Conference (WWW'03)*, Budapest, Hungary, 2003.
- [57] Liu, H.: "MontyLingua: An end-to-end natural language processor with common sense", 2004. Available at: web.media.mit.edu/~hugo/montylingua.
- [58] Luhn, H.P.: "A Statistical Approach to Mechanized Encoding and Searching of Literacy Information". *IBM Journal of Research and Development* 2. pp. 159-165, 1957.
- [59] Maedche, A.; Staab, S.: "Measuring Similarity between Ontologies", In: *Proc. Of the European Conference on Knowledge Acquisition and Management - EKAW-Madrid*, Spain, October 1-4, 2002. LNCS/LNAI 2473, Springer, 2002, pp. 251-263.

- [60] Mandala, R.; Tokunaga, T.; Tanaka, H.; Okumura, A.; Satoh, K.: “Ad Hoc Retrieval Experiments using Wordnet and Automatically constructed thesauri”, TREC98, pp. 414-419, 1998.
- [61] Minker, J.; Wilson, G. G.A.; Zimmerman, B.: “An evaluation of query expansion by the addition of clustered terms for a document retrieval system”. *Information Storage and Retrieval*, 8(6):329–348, 1972.
- [62] Mladenic, D.; Grobelnik, M.: “Feature selection for classification based on text hierarchy”. In *Working Notes of Learning from Text and the Web, Conf. Automated Learning and Discovery (CONALD-98)*. Carnegie Mellon Univ., Pittsburgh, 1998.
- [63] OCLC Forest Press. Introduction to the dewey decimal classification. <http://www.oclc.org/dewey/about/default.htm>
- [64] Park, Y. C.; Han, Y. S.; Choi, K.S.: “Automatic Thesaurus Construction using Bayesian Networks”, in *Proceedings of the 1995 International Conference on Information and Knowledge Management*, pp. 212-217, Baltimore, Maryland, USA, 1995.
- [65] Peat, H. J.; Willett, P.: “The limitations of term co-occurrence data for query expansion in document retrieval systems”. *Journal of the American Society for Information Science*, 42(5):378--383, 1991.
- [66] The PennTree Project <http://www.cis.upenn.edu/~treebank/home.html>
- [67] Qiu, Y.; Frei, H.P.: “Concept based query expansion”. In *Proceedings of ACM SIGIR International Conference on Research and Development in Information Retrieval*, ACM Press, pp. 160-170, 1993.
- [68] Quinlan, J.R.: “Constructing Decision Tree in C4.5: Programs for Machine Learning”, pp. 17-26, Morgan Kaufman Publishers, 1993.
- [69] The Reuters Text Collection
<http://www.daviddlewis.com/resources/testcollections/reuters21578/>
- [70] Van Rijsbergen, C.J.: “Information Retrieval”, 2nd edition, London, Butterworths, 1979.
- [71] Van Rijsbergen, C.J.; Harper, D.J.; Porter, M.F.: “The selection of good search terms, *Information Processing & Management*”, pp.77-91, 1981.
- [72] Ruge, G.: "Experiments on Linguistically Based Term Associations", *Information Processing & management*, 28(3), pp. 317-332, 1992.

- [73] Salton, G.: "Automatic Information Organization and Retrieval". McGraw-Hill, 1968.
- [74] Salton, G.; Buckley, C.: "Term-weighting approaches in automatic text retrieval", *Information Processing and Management: an International Journal*, v.24 n.5, p.513-523, 1988
- [75] Salton, G.; Yang, C.S.; Yu, C.T.: "A theory of term importance in automatic text analysis". *Journal of the American Society for Information Science*, 26(1):33--44, Jan-Feb 1975.
- [76] Sanderson, M.; Croft, W.B.: "Deriving concept hierarchies from text", in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, p.206-213, August 15-19, 1999, Berkeley, California, United States.
- [77] Smeaton, A.F.; Van Rijsbergen C.J.: "The retrieval effects of query expansion on a feedback document retrieval system". *The Computer Journal*, 26(3):239--246, 1983.
- [78] Tan, P.N.; Kumar, V.; Srivastava, J.: "Selecting the right interestingness measure for association patterns", in *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp 32-41, 2002.
- [79] The Topic Detection and Tracking (TDT) project
<http://www ldc.upenn.edu/Projects/TDT/>
- [80] Terra, E; Clarke, C. L. A.: "Frequency Estimates for Statistical Word Similarity Measures". *HLT/NAACL 2003*, Edmonton, Alberta, 2003. 37/162.
- [81] The Text REtrieval Conference (TREC) <http://trec.nist.gov/data.html>
- [82] Turney, P.D.: "Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL". In *Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001)*, pp. 491-502, 2001.
- [83] Turney, P.D.: "Extraction of Keyphrases from Text: Evaluation of Four Algorithms". *Technical Report NRC 41550*, National Research Council of Canada, 1997.
- [84] Wen, J.-R.; Nie, J.-Y.; Zhang, H.-J.: "Clustering User Queries of a Search Engine". *WWW10*, May 1-5, 2001, Hong Kong.
- [85] Wulfekuhler, M.R.; Punch, W.F.: "Finding salient features for personal web page categories". In *Proceedings of the 6 international World Wide Web conference*, 1997.

- [86] Xu, J., Croft, W.B.: “Improving the effectiveness of information retrieval with local context analysis”. *ACM Transactions on Information Systems*, 18(1): 79-112, 2000.
- [87] Yarowsky, D.: “Word-sense disambiguation using statistical models of Roget’s categories trained on large corpora”. In *Proceedings of COLING-92*. pp 454-460, Nantes, France, 1992.
- [88] Yang, Y.; Pedesen, J.O.: “A Comparative Study on Feature Selection in Text Categorization”, in *Proc. of the 14th International Conference on Machine Learning*, pp 412-420, 1997.

CURRICULUM VITAE

Libo Chen

Education

2002 - 2006	Ph.D. candidate in Computer Science, Darmstadt University of Technology, Darmstadt, Germany
1998 - 2001	Master of Computer Science and Economics (Diplom Wirtschaftsinformatik), Darmstadt University of Technology, Darmstadt, Germany
1991 - 1996	Bachelor of Management Information System, Tsinghua University, Beijing, China
1988 - 1991	Senior Middle School of Peking University, Beijing, China
1985 - 1988	Junior Middle School of Huayuancun, Beijing, China

Working Experiences

2002 -	Regular employment of Fraunhofer IPSI, Darmstadt
1996 - 1997	Regular employment of the Automatic Institute, Chinese Academy of Sciences